EXHIBIT A

IN THE UNITED STATES DISTRICT COURT
FOR THE DISTRICT OF MASSACHUSETTS

NUANCE COMMUNICATIONS, INC., Plaintiff and Counterclaim Defendant, v. OMILIA NATURAL LANGUAGE SOLUTIONS, LTD., Defendant and Counterclaim Plaintiff.	Case No. 1:19-CV-11438-PBS

OMILIA NATURAL LANGUAGE SOLUTIONS, LTD.'S SECOND SUPPLEMENTAL PRELIMINARY NON-INFRINGEMENT AND INVALIDITY CONTENTIONS PURSUANT TO LOCAL RULE 16(d)(4)

Omilia Natural Language Solutions, Ltd. ("Omilia NLS"), pursuant to Local Rule 16.6(d)(4)(D) and (E) and the Court's Scheduling Order (ECF No. 55), submits its Second Supplemental Non-Infringement Contentions and Invalidity Claim Charts concerning U.S. Patent Nos. 6,999,925 (the "'925 Patent") and 8,532,993 (the "'993 Patent") (collectively the "asserted patents").

For at least the reasons discussed herein, the asserted patents are not infringed and invalid. Omilia NLS reserves the right to amend, supplement, and/or materially modify these Non-Infringement and Invalidity Contentions pursuant to the Federal Rules of Civil Procedure, the Court's Local Rules, and/or any other order or schedule entered by the Court.

OMILIA NLS' SUPPLEMENTAL NON-INFRINGEMENT AND INVALIDITY CONTENTIONS (JUNE 9, 2020)

I. PRELIMINARY STATEMENT

1. At the October 30, 2019 case management conference, the Court ordered Nuance to narrow the asserted patents in this matter. Following the Court's order, Nuance selected the

’925 Patent and the ’993 Patent and elected to stay all other asserted patents. Omilia NLS serves these contentions as to the ’925 Patent and ’993 Patent and reserves all rights to serve further contentions with regard to the remaining patents at an appropriate time as determined by the parties and the Court.

2. On December 6, 2019, Nuance served its preliminary infringement contentions for the ’925 and ’993 Patents. Nuance asserted all claims, a total of 49, across the two patents. Nuance’s contentions fail to sufficiently describe Nuance’s infringement theories or indicate how it believes Omilia NLS’ products infringe the asserted claims. As a result, Omilia NLS has been prejudiced in understanding and responding to Nuance’s allegations. Omilia NLS further reserves the right to respond to these allegations if Nuance is permitted to provide further arguments or evidence.

3. These supplemental contentions are limited to Omilia NLS’ products accused in this litigation that are sold, offered for sale or available to users in the United States.

4. These non-infringement and invalidity contentions are based on Omilia NLS’ current knowledge, understanding, and belief as to the facts and information available as of the date of these contentions. Discovery is ongoing and Omilia NLS has not completed its investigation, discovery, or analysis of information related to this action. Omilia NLS reserves the right to amend, modify and/or supplement its non-infringement and invalidity contentions.

5. Nothing stated herein is or shall be treated as an admission or suggestion that Omilia NLS agrees with Nuance regarding either the scope of any of the asserted claims or the claim construction positions advanced expressly or implicitly by Nuance’s Preliminary Infringement Contentions or in any other pleading, discovery request or response, or any written or verbal communication with Omilia NLS. Additionally, nothing in these Non-Infringement and

Invalidity Contentions shall be treated as an admission that any of Omilia NLS' accused products meet any limitation of the asserted claims. The disclosures herein are not, and should not be construed as a statement that no other persons have discoverable information, that no other documents, data compilations, and/or tangible things exist that Omilia NLS may use to support their claims or defenses, or that no other legal theories or factual bases will be pursued.

6. This preliminary disclosure is directed to non-infringement and invalidity issues only and does not address unenforceability, claim construction, waiver, estoppel, laches (prosecution laches), failure to mark, limitations on damages under 35 U.S.C. § 286 and 28 U.S.C. § 1498, or any other defense precluding the enforcement of the asserted patents, or barring or limiting damages or injunctive relief. Omilia NLS reserves all rights with respect to such issues. In addition, Omilia NLS has not deposed any of the inventors or the prosecuting attorneys of the asserted patents. Omilia NLS reserves the right to raise any issues or defenses that come to light during discovery, including defenses concerning inventorship and/or inequitable conduct.

7. Omilia NLS does not assign any particular claim construction or scope to the claims. The timing and manner of claim construction contentions are governed by Local Rule 16(e) and the Court's Scheduling Order (ECF No. 55). Omilia NLS reserves all rights to make any claim construction arguments as permitted under the rules and recognizes final claim construction and scope will be determined by the Court.

8. Omilia NLS is providing these non-infringement and invalidity contentions prior to any claim construction ruling by the Court. Any non-infringement analysis depends, ultimately, upon claim construction, which is a question of law reserved for the Court. Omilia NLS reserves the right to amend, supplement, or materially modify its non-infringement and invalidity contentions after the claims have been construed by the Court. Omilia NLS also reserves the right

to amend, supplement, or materially modify its non-infringement and invalidity contentions based on any claim construction positions that Nuance may take in this case. Moreover, since claim construction has yet to occur, Omilia NLS has been forced to make assumptions about the meaning of certain terms. The assumptions, or the lack thereof, made by Omilia NLS for the purposes of these non-infringement and invalidity contentions do not necessarily convey Omilia NLS' positions as to the meaning or validity of claim terms and cannot be used as evidence of Omilia NLS' views in future proceedings involving claim construction or invalidity of the claims, including under 35 U.S.C. § 112.

9. Omilia NLS is providing these supplemental non-infringement and invalidity contentions prior to any expert discovery. Omilia NLS reserves the right to amend, supplement, or materially modify its non-infringement contentions in view of expert discovery.

II. NON-INFRINGEMENT CONTENTIONS

A. The '925 Patent

Omilia NLS does not infringe the asserted claims of the '925 patent for at least the following reasons:

1. Nuance Cannot Demonstrate that Omilia NLS' Products Infringe

Nuance's infringement contentions fail to provide a sufficient basis for infringement either literally or under the doctrine of equivalents. For example, Nuance fails to demonstrate that each of the steps in method claims 1-13 and 27-29 is practiced in the United States. Nuance also fails to demonstrate that a "second speech recognizer" is generated and how it is generated as required by all the claims, that the alleged "second speech recognizer" has a "second decision network" or "second phonetic contexts," that the "second decision network is not fixed by the number of nodes in the first decision network," or that "re-estimating comprises partitioning said training data using said first decision network of said fair speech recognizer." Nuance provides no allegations or

evidence to support its claims that these elements are literally satisfied. *See e.g.,* Nuance's December 6, 2019 Infringement Contentions, Exhibit A at A-15. Nuance's allegations with respect to the dependent claims are equally deficient and fail to provide a meaningful basis to understand or evaluate their claims of infringement.

Nuance provides no allegations whatsoever concerning how Omilia NLS' products infringe under the doctrine of equivalents. Moreover, as explained below, Omilia NLS' products do not infringe under the doctrine of equivalents as the differences between those products and the '925 patent are substantial. For the reasons set forth below, the Omilia NLS' products do not substantially perform the same function in substantially the same way to obtain the same result. *See infra.* Nuance is estopped from claiming infringement as an equivalent to the extent Nuance's theories ensnare prior art or seek coverage over designs disclaimed during the prosecution of the '925 patent. As Nuance has not provided any potential equivalents, Nuance may be foreclosed from asserting infringement analysis to the extent Nuance's arguments are ensnared by the prior art.

Because Nuance's contentions remain deficient, Omilia NLS reserves the right to further supplement its contentions to account for any additional arguments or evidence presented by Nuance.

2. [redacted]

Independent claim 27 and dependent claims 28 and 29 claim a "second speech recognizer," which "is a multi-lingual speech recognizer." *See, e.g.*, the '925 Patent at 14:22-23. Omilia NLS does not infringe claims 27-29 of the '925 Patent because [redacted]

3. [redacted]

Independent claim 27 and dependent claims 28 and 29 require a "second domain comprising a second acoustic model." *See, e.g.*, the '925 Patent at 14:17-18. [redacted]

4. [redacted]

Claims 1-26 of the '925 Patent claim a method (claims 1-13) or a machine-readable storage (claims 14-26) that includes "automatically generating from a first speech recognizer a second speech recognizer." *See, e.g.*, the '925 Patent at 10:42-43. Omilia NLS' software does not automatically generate a second speech recognizer from a first speech recognizer. [redacted]

Accordingly, Omilia NLS' products do not infringe claims 1-26. For the same reasons Omilia NLS' products also do not infringe claims 27-29 of the '925 Patent directed to "a computerized method of generating a second speech recognizer," [redacted]

5. [redacted]

All of the asserted claims of the ’925 Patent require a “second speech recognizer” or “second acoustic model” generated from “said first decision network and said corresponding first phonetic contexts based on domain-specific training data.” *See, e.g.*, the ’925 Patent at 10:51-53. Nuance’s infringement contentions do not sufficiently identify a second speech recognizer such that Omilia NLS infringes the claims of the ’925 patent. [redacted] Therefore, Omilia NLS’ products do not infringe any claims of the ’925 Patent.

[redacted] As the ’925 Patent explains, “re-estimation” means “a complete recalculation of the decision network and its corresponding phonetic contexts based on the general speech recognizer decision network.” The ’925 Patent at 7:7-11. This re-estimation requires the “adaptation of the recognizer’s sub-word inventory to a special domain.” *Id.* at 6:66-7:2. [redacted]

Additionally, [redacted] as claimed in claims 1, 3-14, 16-26 of the ’925 Patent. *See, e.g.*, the ’925 Patent at 10:59-61. The ’925 Patent explains, “one obtains partitioning of the adaption data that already utilizes the phonetic context information of the much larger and more general training corpus of the base system.” *Id.* at 7:58-61. [redacted]

6. [redacted]

All of the asserted claims of the ’925 Patent require that the “second speech recognizer” or “acoustic model” has (1) “a second decision network”, and (2) “corresponding second phonetic contexts.” *See, e.g.*, the ’925 Patent at 10:49-51. [redacted] Therefore, Omilia NLS’ products do not infringe any claims of the ’925 Patent.

7. [redacted]

Claims 1-26 of the ’925 Patent require that “the number of nodes in the second decision network is not fixed by the number of nodes in the first decision network.” *See, e.g.,* the ’925 Patent at 10:56-59. Independent claims 2 and 15 further require “adding a node to the second decision network for the identified context independent of other generating step operations.” *See, e.g.*, the ’925 Patent at 11:17-19. [redacted] Accordingly, Omilia NLS does not infringe claims 1-26 of the ’925 Patent.

8. Nuance Has Not Sufficiently Identified the Claimed Generation Based On “Domain-Specific Training Data”

Claims 1-29 of the ’925 patent require that the second speech recognizer use “domain-specific training data” to train the second speech recognizer. *See, e.g.*, the ’925 patent at 10:52–53; 11:4–6; 12:14–16; 12:38–40; 14:13–14. Similarly, claims 2 and 15 specify that the domain-

specific training data is "of a limited amount." *Id.* at 11:12-14; 12:46-47. In its infringement contentions, Nuance has not sufficiently identified the claimed domain-specific training data, nor explained how the domain-specific training data is used to generate the alleged second speech recognizer.

9. [redacted]

Claims 1-13 and 27-29 of the '925 Patent are directed to methods of "generating" "a second speech recognizer" and include the step of "generating a second acoustic model." *See, e.g.*, the '925 Patent at 10:42-49. [redacted] required by claims 1-13 and 27-29, such work is performed by Omilia NLS' personnel located outside of the United States. Likewise, to the extent the step is performed at all, the [redacted] step of claims 27-29 is performed by Omilia NLS' personnel located outside of the United States. [redacted] and could not be alleged to perform all elements of the methods of claims 1-13 and 27-29 in the United States. For these reasons, Omilia NLS does not infringe claims 1-13 and 27-29 of the '925 Patent.

10. Indirect Infringement

Nuance alleges that Omilia NLS induced and/or contributed to its customers' infringement of the '925 Patent. [redacted] Accordingly, none of Omilia NLS' customers directly infringe the claims of the '925 Patent. Omilia NLS does not induce or contribute to infringement of claims 1-13 and 27-29 of the '925 Patent because it does not enable its customers to perform the claimed

method steps. Additionally, Omilia NLS does not induce or contribute to infringement of claims 14-26 of the '925 Patent [redacted]

B. The '993 Patent

Omilia NLS does not infringe the asserted claims of the '993 patent for at least the following reasons:

1. Nuance Cannot Demonstrate that Omilia NLS' Products Infringe

Nuance's infringement contentions fail to allege a sufficient basis for infringement either literally or under the doctrine of equivalents. Nuance fails to demonstrate that each of the steps in method claims 1-8 is practiced in the United States or that systems (claims 9-16), or computer-readable storage devices (claims 17-20), having stored instructions to perform the recited methods are sold or used in the U.S. Nuance further fails to demonstrate how Omilia NLS' products "approximat[e] transcribed speech using a phonemic transcription dataset associated with a speaker, to yield a language model" and "where the phonemic transcription dataset is based on a pronunciation model of the speaker." Nuance fails to demonstrate how Omilia NLS' products "incorporate, into the language model, pronunciation probabilities associated with respective unique labels for each different pronunciation of a word," where there are "unique labels for a most frequent word" and that unique "label indicates a special status in the language model." Nuance further provides no evidence of how or when a language model with the incorporated attributes "recognizes an utterance using the language model." Nuance provides no allegations or evidence to support its claims that these elements are literally satisfied. *See e.g.* Nuance's December 6, 2019 Infringement Contentions, Exhibit B at B-3-6.[1] Nuance's allegations with

[1] These deficiencies are also present in all of the dependent claims of the '993 Patent. *See, e.g.*, Nuance's December 6, 2019 Infringement Contentions, Exhibit B at B-7-17.

respect to the dependent claims are equally deficient and fail to provide a meaningful basis to understand or evaluate their claims of infringement.

Likewise, Nuance provides no allegations concerning how Omilia NLS' products infringe under the doctrine of equivalents. Omilia NLS' products do not infringe under the doctrine of equivalents as the difference between those products and the '993 Patent are substantial. Omilia NLS' products do not substantially perform the same function in substantially the same way to obtain the same result as the claims of the '993 Patent. *See infra.* Nuance is estopped from claiming infringement as an equivalent to the extent Nuance's theories ensnare prior art or seek coverage over designs disclaimed during the prosecution of the '993 Patent. As Nuance has not alleged any equivalents in its Invalidity Contentions, Nuance may be foreclosed from asserting infringement analysis to the extent Nuance's arguments are ensnared by the prior art.

Because Nuance's contentions remain deficient, Omilia NLS reserves the right to further supplement its contentions to account for any additional arguments or evidence presented by Nuance.

2. [redacted]

All claims of the '993 Patent require "incorporating, into the language model, pronunciation probabilities associated with respective unique labels for each different pronunciation of a word." *See, e.g.*, the '993 Patent at 12:19-21. [redacted] Accordingly, Omilia NLS' products

do not infringe any claims of the ’993 Patent.

3. [redacted]

All claims of the ’993 Patent require “unique labels for each different pronunciation of a word” wherein a “unique label for a most frequent word indicates a special status.” *See. e.g.*, the ’993 Patent at 12:20-23. [redacted] Accordingly, Omilia NLS’ products do not infringe any claims of the ’993 Patent.

4. [redacted]

All claims of the ’993 Patent require a “transcription dataset associated with a speaker, to yield a language model, where the phonemic transcription dataset is based on a pronunciation model of the speaker.” *See, e.g.*, the ’993 Patent at 12:15-18. [redacted] Accordingly, Omilia NLS’ products do not infringe any claims of the ’993 Patent.

5. [redacted]

All claims of the ’993 Patent require “after incorporating the pronunciation probabilities into the language model, recognizing an utterance using the language model.” *See, e.g.*, the ’993 Patent at 12:14-26. [redacted] Accordingly, Omilia NLS’ products do not infringe any claims of the ’993 Patent.

6. Omilia's Speech Transcription Operations are Performed Outside the United States

Claims 1-8 of the '993 Patent include the steps [redacted] To the extent Omilia NLS activities can be construed as the performance of any of the steps of the claimed methods, the steps are performed at least in part by Omilia NLS' personnel located outside of the United States. To the extent any such steps that are alleged to be evidence of infringement are performed at all, the [redacted] steps required by claims 1-8 are performed in part outside of the United States. Therefore, it is not possible that all elements of claims 1-8 are performed by Omilia in the United States.

For these reasons, Omilia NLS does not infringe claims 1-8 of the '993 Patent.

7. Indirect Infringement

Nuance asserts that Omilia NLS induced and/or contributed to its customers' infringement of the '993 Patent. [redacted] No Omilia NLS' customer directly infringes any of the claims of the '993 Patent. Accordingly, Omilia NLS does not induce or contribute to infringement of any claims of the '993 Patent.

C. Omilia NLS Does Not Induce or Contribute to Infringement of the '925 and '993 Patents

Omilia NLS does not induce or contribute to infringement of the '925 Patent or '993 Patent because there is no direct infringement by Omilia NLS' customers. Therefore, Omilia NLS is not liable for contributory or induced infringement.

Nuance has failed to allege any facts that would give rise to induced infringement. Induced infringement requires both knowledge of the patent and specific intent to cause infringement. *See DSU Med. Corp. v. JMS Co.*, 471 F.3d 1293, 1306 (Fed. Cir. 2006) ("[T]he 'alleged infringer must be shown to have *knowingly* induced infringement,' not merely knowingly induced the *acts* that constitute direct infringement.") (emphasis in original) (citations omitted) (quoting *Manville Sales Corp. v. Paramount Sys.*, 917 F.2d 544, 553 (Fed. Cir. 1990)); *accord Commil USA, LLC v. Cisco Sys.*, 135 S. Ct. 1920, 1928 (2015) ("[Induced infringement] requires proof the defendant knew the acts were infringing."). Nuance has failed to properly allege induced infringement by alleging only knowledge of the asserted patents and conclusory allegations that Omilia NLS knows the Accused IVR Platform operates in a manner that infringes the asserted patents. Such conclusory allegations are insufficient and Nuance has failed to provide evidence that Omilia NLS had the specific intent to induce acts that constitute infringement. *See DSU Med.*, 471 F.3d at 1306. Nuance cannot show that Omilia NLS had knowledge or specific intent to induce infringement at least because Omilia NLS has a good-faith belief that the asserted patents are not infringed.

Nuance has similarly failed to allege any facts that would give rise to contributory infringement. Contributory infringement requires the sale or offer of "a component of a patented machine, manufacture, combination or composition, or a material or apparatus for use in practicing a patented process, constituting a material part of the invention . . . ." *See* 35 U.S.C. § 271(c). Nuance provides only conclusory allegations that the entire Accused IVR Platform is a material part of the purported inventions. Such conclusory allegations do not and cannot demonstrate that a component sold by Omilia NLS "constitutes a material part" of the purported inventions covered by the asserted patents.

D. Omilia NLS Does Not Willfully Infringe the Asserted Patents

Nuance has failed to meet its burden to prove willfulness. Nuance's Infringement

Contentions have alleged nothing more than Omilia NLS' knowledge of the asserted patent, mere knowledge of the patents is insufficient for a finding of willfulness. *See, e.g.*, *Trs. of Bos. Univ. v. Everlight Elecs. Co., Ltd.*, 212 F. Supp. 3d 254, 256-57 (D. Mass. 2016) (A court may not "award enhanced damages simply because the evidence shows that the infringer knew about the patent *and nothing more*.") (citing *Halo Elecs., Inc. v. Pulse Elecs., Inc.*, 136 S. Ct. 1923, 1936 (2016)) (emphasis in original).

Nuance has similarly failed to show that enhanced damages are appropriate even if the Court were to find willfulness because such punishment should generally be reserved for egregious cases typified by willful misconduct. *Koninklijke Philips N.V. v. Zoll Med. Corp.*, 257 F. Supp. 3d 159, 163 (D. Mass. 2017) ("Plaintiffs have not met their burden to show that defendant's conduct was so malicious that a finding of willful infringement is warranted in this case"); *see also Brigham & Women's Hosp., Inc. v. Perrigo Co.*, 251 F. Supp. 3d 285, 293 (D. Mass. 2017) (for enhanced damages under *Halo*, infringement must be willful and egregious); *Everlight Elecs.*, 212 F. Supp. 3d at 258 ("Assuming without deciding that the jury's verdict, based on the subjective prong of the now-overruled *Seagate* test, is sufficient to find subjective willfulness, the Court still finds, in its discretion, that the defendants' conduct did not rise to the level of egregiousness meriting an award of enhanced damages.")

III. INVALIDITY CONTENTIONS

A. The '925 Patent

As described in Exhibit A, at least the prior art references below render the claims of the '925 Patent invalid.

- United States Patent No. 6,912,499, Sabourin et al., filed August 31, 1999 ("Sabourin").
- United States Patent No. 6,336,108, Thiesson et al., filed December 23, 1998 ("Thiesson").
- United States Patent No. 7,216,079, Barnard et al., filed November 2, 1999

("Barnard").

- United States Patent Publication No. 2008-0147404, Liu et al., priority to May 15, 2000 ("Liu").
- United States Patent No. 6,151,574, Lee et al., filed September 8, 1998 ("Lee").
- Duchateau et al., *A Novel Node Splitting Criterion in Decision Tree Construction for Semi-Continuous HMMS*, 5th European Conference on Speech Communication and Technology EUROSPEECH '97 (1997) ("Duchateau").
- Schultz, et al., *Japanese LVCSR on the Spontaneous Scheduling Task with JANUS-3*, Eurospeech, Rhodes 1997 ("Schultz").
- M. Finke, et al., *Wide Context Acoustic Modeling in Read vs. Spontaneous Speech*, Proc. Of ICASSP, Munich 1997 ("Finke").
- V. Fischer, et al., *Speaker-Independent Upfront Dialect Adaptation in a Large Vocabulary Continuous Speech Recognizer*, 5th International Conference on Spoken Language Processing, Sydney, Australia 1998 ("Fischer").
- United States Patent No. 6,324,510, Waibel et al., filed November 6, 1998 ("Waibel").
- United States Patent No. 6,789,061, Fischer et al., filed August 14, 2000 (the "'061 Patent").

For example, the '925 patent is anticipated under § 102 or rendered obvious under § 103 in light of one or more of the below combinations of prior art.

1. Thiesson anticipates claim 1 of the '925 Patent and renders obvious claims 2-29 alone or in combination with Liu, Lee and Duchateau.
2. Sabourin renders obvious, alone or in combination with Thiesson, Liu, Lee, and Duchateau, all claims of the '925 Patent.
3. Barnard renders obvious, alone or in combination with Thiesson, Liu, Lee, and Duchateau, all claims of the '925 Patent.
4. Schultz renders obvious, alone or in combination with Finke, Fischer, and Sabourin, all claims of the '925 Patent.
5. Waibel in view of Sabourin renders obvious claims 27-29 of the '925 Patent.
6. The '061 Patent anticipates claims 1, 13, 14, and 26 of the '925 Patent.

Additionally, the claims of the '925 Patent are invalid for Obvious-Type Double Patenting over the '061 Patent". Specifically:

1. Claim 1 is invalid over claim 6 of the ’061 Patent.
2. Claims 3-11, and 13 are invalid over claims 3 and 6 of the ’061 Patent.
3. Claims 14, 16-24, and 26 are invalid over claims 12 and 15 of the ’061 Patent.
4. Claim 2 is invalid for double patenting over claim 6 of the ’061 Patent in view of Schultz.
5. Claim 15 is invalid for double patenting over claim 15 of the ’061 patent in view of Schultz.
6. Claims 12 and 27-29 are invalid over claim 6 of the ’061 Patent in view of Schultz and Sabourin.
7. Claim 25 is invalid over claim 15 of the ’061 Patent in view of Schultz and Sabourin.

The prior art references all relate to methods of generating speech recognizers. A person of ordinary skill in the art would be motivated to combine the references with known elements of other speech recognizers to create better more efficient speech recognizers. A person of ordinary skill in the art would be motivated to combine the references. All of the references are directed to building an improved speech recognizer. It would have been obvious to implement the teachings of the references in a way corresponding to the ’925 patent. Combining the references in the above manner would have employed a known technique such as disclosed in the ’925 patent. As such, a person of ordinary skill would expect the combinations above would yield predictable and successful results. Nuance has not yet indicated what secondary considerations of non-obvious it plans to rely on. Omilia NLS reserves the right to amend or supplement its contentions if Nuance elects to pursue secondary considerations.

In addition, the claims of the ’925 Patent are directed to an abstract idea because the claims

relate to ineligible subject matter under 35 U.S.C. § 101. The performance of a mathematical formula and abstract idea by a machine is not "an inventive concept sufficient to transform the claimed abstract idea into a patent eligible application," and thus, renders the claims unpatentable subject matter. *See Alice Corp. Pty. Ltd v. CLS Bank Int'l*, 134 S. Ct. 2347, 2357 (2014). For example, the claims of the '925 Patent generally concern the abstract idea of generating a second speech recognizer by modifying a first speech recognizer for a different type of speech (or language) based on training information from that different speech. The claims are, thus, directed to abstract mental steps that can be performed without a computer. The claims do not recite a method or machine readable storage that improves the functionality of a computer or addresses a computer-specific issue. Additionally, the limitations and combination of limitations do not provide an inventive contribution. Rather, the limitations recite either mental steps, steps that could be done by a human without aid of a computer, or well-understood, conventional or routine activities performed on a generic processor. The scope of the claims preempt the entire field of use relating to modifying existing speech detection techniques using training data from another language or field. Thus, the claims of the '925 Patent fail to transform the claimed abstract idea into a patent-eligible subject matter.

In addition, the '925 Patent claims are invalid under 35 U.S.C. § 112(a) because they fail to provide sufficient written description and enablement for what is claimed. For example:

- The '925 Patent fails to provide sufficient disclosure or enablement for "automatically generating" the second speech recognizer.
- To the extent the claims of the '925 Patent encompass "re-estimation" methods beyond the limited method of "re-estimation" as described and defined in the '925 Patent (*see e.g.*, 6:66-7:16), the claims are invalid for lack of written description under 35 U.S.C. § 112.

- The ’925 Patent fails to enable under 35 U.S.C. § 112 the generation of a “second speech recognizer” as claimed or how to “add[] a node to the second decision network for the identified context independent of other generating step operations”.
- To the extent the claims of the ’925 Patent encompass subject matter that extends beyond the “generating” as described and defined in the ’925 Patent (*see e.g.*, 6:66-7:16), the claims are invalid under 35 U.S.C. § 112 for lack of written description and enablement.
- The ’925 Patent fails to describe or enable what it means to “partition training data using said first decision network of said first speech recognizer,” and are therefore invalid under 35 U.S.C. § 112.

In addition, the ’925 Patent asserted claims are invalid under 35 U.S.C. § 112(b) because they are indefinite and fail to define the scope of what is claimed. The term “the number of nodes in the second decision network is not fixed by the number of nodes in the first decision network” (claims 1-26) fails to define the scope of the asserted claims, to the extent this limitation is understood to cover the second decision network with the same nodes (as well as number) as the first decision network. To the extent the claims are understood to include any “generating” steps other than those described in the specification, the term “generating… a second speech recognizer” is indefinite.

These invalidity arguments, including the prior art references and their use described above (and in Exhibit A), are merely representative and this list is not meant to limit Omilia NLS’ invalidity arguments in this case. Omilia NLS reserves the right to supplement these contentions as the case proceeds, in particular in light of the Court’s claim construction order as well as in connection with expert discovery.

B. The ’993 Patent

As described in Exhibit B, at least the prior art references below render the claims of the ’993 Patent invalid.

- Mirjam Wester, *Pronunciation Variation Modeling for Dutch Automatic Speech Recognition*, 2002 (“Wester”) and exemplary publications incorporated and discussed therein:
  - Kessens et al., *Improving the performance of a Dutch CSR by modeling within-word and cross-word pronunciation variation*, Speech Communication 29 (1999) 193-07 (incorporated as pp. 49-65 in Wester) (“Kessens”)
  - Wester, *Pronunciation Modeling for ASR – Knowledge-based and Data-derived Methods*, 2001 (incorporated as pp. 97-122 in Wester) (“Wester2”)
- Steinbiss et al., *The Philips research system for large-vocabulary continuous-speech recognition*, Proc. of 3rd European Conference on Speech Communication and Technology EUROSPEECH ‘93, 2125-28 (1993) (“Steinbiss”)
- Bahl et al., *Context Dependent Modeling of Phones in Continuous Speech Using Decision Trees*, HLT ‘91 Proceedings of the workshop on Speech and Natural Language (1991) 264-69 (“Bahl”)

For example, the ’993 patent is anticipated under § 102 or rendered obvious under § 103 in light of one or more of the below combinations of prior art.

1. Wester anticipates each and every claim of the ’993 Patent.
2. Wester renders obvious, alone or in combination with Bahl, claims 4, 12, and 20 of the ’993 Patent.
3. Wester renders obvious, alone or in combination with Steinbiss, claims 9 through 20 of the ’993 Patent.

The references all relate to methods of generating speech recognizers. A person of ordinary skill in the art would be motivated to combine the references with known elements of other speech recognizers to create more efficient and effective speech recognizers. A person of ordinary skill in the art would be motivated to combine the references. All of the references are directed to improved pronunciation models in speech recognition systems. It would have been obvious to implement the teachings of the references in a way corresponding to the ’993 patent. Combining

the references in the above manner would have employed a known technique such as disclosed in the ’993 patent. As such, a person of ordinary skill would expect the combinations above would yield predictable and successful results. Nuance has not yet indicated what secondary considerations of non-obvious it plans to rely on. Omilia reserves the right to amend or supplement its contentions if Nuance elects to pursue secondary considerations.

The claims of the ’993 Patent are directed to an abstract idea and claim ineligible subject matter under 35 U.S.C. § 101. The performance of a mathematical formula or abstract idea by a machine is not “an inventive concept sufficient to transform the claimed abstract idea into a patent eligible application,” and thus, renders the claims unpatentable subject matter. *See Alice Corp. Pty. Ltd,* 134 S. Ct. at 2357. The claims are directed to abstract mental steps that can be performed without a computer. The recited claims do not recite a method or system that improves the functionality of a computer or addresses a computer-specific issue. Additionally, the limitations and combination of limitations do not provide an inventive contribution. Rather, the limitations recite either mental steps, steps that could be done by a human without aid of a computer, or well-understood, routine, or conventional activities performed on a generic processor. The scope of the asserted claims preempt the entire field of use pronunciation probabilities to modifying existing speech detection techniques. Thus, the claims of the ’993 Patent fail to transform the claimed abstract idea into a patent-eligible subject matter.

The ’993 Patent asserted claims are invalid under 35 U.S.C. § 112(a) because they fail to provide sufficient written description and enablement for what is claimed.

- The ’993 Patent fails to provide sufficient disclosure or enablement for the “incorporating” limitation present in all claims. There is insufficient disclosure or enablement for “incorporating into a language model” as well as “incorporating . . . probabilities associated

with respective unique labels for each different pronunciation of a word."

- The limitation "unique label for a most frequent word indicates a special status" does not refer to an integer representing which pronunciation is associated with the lexical entry. There is no written description or enablement under 35 U.S.C. § 112 for all claims of the '993 Patent and the claims are, therefore, invalid.

The asserted claims of the '993 Patent are invalid under 35 U.S.C. § 112(b) because they are indefinite and fail to define the scope of what is claimed. The following terms fail to define the scope of the asserted claims:

- "**approximating** transcribed speech" (all claims)
- "**incorporating**, into the language model" (all claims)
- "most frequent word" (all claims)
- "recognizing an utterance" (all claims)
- "modeling pronunciation dependencies across word boundaries" (claims 3, 11, 19)
- "contextual dependency" (claims 4, 12, 20)
- "consistency in pronunciation probabilities" (claims 4, 12, 20)
- "throughout the utterance" (claims 4, 12, 20)
- "a set of **most frequent words** each with more than one **pronunciation alternative**" (claim 7)
- "pronunciation alternative" (claim 8)

These invalidity arguments, including the prior art references and their use described above (and in Exhibit B), are merely representative and this list is not meant to limit Omilia NLS' invalidity arguments in this case. Omilia NLS reserves the right to supplement these contentions as the case proceeds, in particular in light of the Court's claim construction order as well as in

connection with expert discovery.

OMILIA NLS' SECOND SUPPLEMENTAL NON-INFRINGEMENT AND INVALIDITY CONTENTIONS (OCTOBER 30, 2020) [2]

IV. SECOND SUPPLEMENTAL NON-INFRINGEMENT CONTENTIONS

In addition to its prior Non-Infringement Contentions, which Omilia NLS hereby reaffirms and reincorporates herein, Omilia NLS further states as follows:

Omilia objects to Nuance's improper and incorrect generalizations of the operation of Omilia's software. In its contentions, Nuance does not show any specific act of infringement for any specific ASR component created, deployed or sold by Omilia NLS, nor does Nuance demonstrate that the process used to generate any specific ASR components is representative of all of Omilia NLS' ASR components that it provides to its customers. Moreover, Nuance has failed to demonstrate what conduct or use, if any, is performed or occurs in the United States or meets any other basis for liability within the United States. Both of these deficiencies are true for all of Nuance's infringement allegations. Instead, Nuance talks about Omilia's software in generalities, without demonstrating an actual and specific model that infringes any of the claims. This deficiency is further compounded by the fact that Nuance's allegations cobble together disparate functionality and software that does not work together or concerns different elements. Nuance likewise cites to legacy or reference software without any indications that they have been used, let alone what model infringes and how there is liability within the United States. It is Nuance's burden to demonstrate that these elements are met and Nuance continues to have no basis to do so.

Contrary to Nuance's arguments, Nuance's inability to establish that Omilia's products infringe the '925 and '993 patents is not because certain discovery has not been provided to

[2] This is when Omilia sought leave of the Court to amend its invalidity contentions.

Nuance, but because Omilia's products operate in a fundamentally different manner than the '925 and '993 patents. Nuance has had and continues to have access to all of the Omilia source code governing the creation and use of its ASR technology as well as ASR components demonstrating the results of that code and the development, use and deployment of those components. Omilia has also made available its entire GitLab repository. Any further discovery is not relevant, duplicative, or non-proportional to the needs of this case. In any case, Omilia has provided or agreed to provide the additional requested materials.

Omilia further objects to Nuance's contentions to the extent that they accuse actions by Omilia's customers taken independently of Omilia. Omilia does not direct or control its customers' performance, nor does it form a joint enterprise with its customers. In addition, Omilia does not condition participation in an activity or the receipt of a benefit upon any actions by its customers.

A. The '925 Patent

On August 24, 2020, Nuance narrowed the asserted '925 patent claims to 1, 14, and 27. Nuance's allegations of infringement are limited to Omilia's alleged use of "transfer learning" to train its ASR technology. Nuance's Response to Omilia's Interrogatory No. 9, Ex. A. Nuance has not alleged infringement of the '925 patent by any other aspects of Omilia's products. Further, Nuance has not identified a single alleged "acoustic model" or "speech recognizer" that was generated using the method claimed in the '925 patent. Moreover, Nuance has failed to demonstrate what conduct or use, if any, is performed or occurs in the United States or meets any other basis for liability within the United States.

Omilia NLS does not infringe the '925 patent claims as the accused Omilia products operate in a fundamentally different manner. For example, Omilia NLS' software does not create a second speech recognizer by re-estimating a first speech recognizer, as required by the claims of

the ’925 patent. Furthermore, nearly every limitation of the asserted claims operates in a manner different than the recited elements of the ’925 patent. As described in Appendix 1, and as illustrated in the documents cited therein, which are incorporated by reference herein, Omilia NLS’ software does not practice, either literally or under the doctrine of equivalents, multiple claim limitations of the asserted claims of the ’925 patent. As Omilia has previously raised with Nuance, Nuance still has no good faith basis to continue to assert these claims given these fundamental differences.

With regard to the doctrine of equivalents, Nuance’s infringement contentions fail to identify any aspect of any accused product that purportedly infringes any claim or any element of any asserted claim under the doctrine of equivalents. *See Intendis GMBH v. Glenmark Pharm. Inc., USA*, 822 F.3d 1355, 1360 (Fed. Cir. 2016) (“[e]ven when an accused product does not meet each and every claim element literally, it may nevertheless be found to infringe the claim ‘if there is ‘equivalence’ ***between the elements of the accused product or process and the claimed elements of the patented invention.***’”). Indeed, Nuance’s infringement contentions do not even identify what specific products allegedly infringe under the doctrine of equivalents. Instead, Nuance’s infringement contentions generically set forth boilerplate and include only conclusory allegations of infringement under the doctrine of equivalents. *See* Nuance’s Response to Omilia’s Interrogatory No. 9, Ex. A. This is insufficient as a matter of law. *See, e.g.*, *Oil-Dri Corp. of Am. v. Nestlé Purina Petcare Co.*, No. 15 C 1067, 2018 WL 1071443, *5 (N.D. Ill. Feb. 26, 2018) (striking doctrine of equivalents arguments where they “do not sufficiently address why the purported aspects of the Accused Products are equivalent and why any differences are insubstantial”). To the extent Nuance is permitted to supplement its infringement contentions with a doctrine of equivalents infringement theory, Omilia reserves the right to respond.

B. The ’993 Patent

On August 24, 2020, Nuance narrowed the asserted ’9993 patent claims to 17 and 19. Further, Nuance has not identified a single alleged “language model” that was generated using the method claimed in the ’993 patent, let alone any computer readable storage media with the instructions for the steps claimed in the ’993 patent. Moreover, Nuance has failed to demonstrate what conduct or use, if any, is performed or occurs in the United States or meets any other basis for liability within the United States.

Omilia NLS does not infringe the ’993 patent claims as the accused Omilia products operate in a fundamentally different manner. For example, Omilia NLS’ software does not use pronunciation probabilities in a dictionary, let alone a language model as required by the claims of the ’993 patent. Furthermore, nearly every limitation of the asserted claims operates in a manner different than the recited elements of the ’993 patent. As described in Appendix 2, and as illustrated in the documents cited therein, which are incorporated by reference herein, Omilia NLS’ software does not practice, either literally or under the doctrine of equivalents, multiple claim limitations of the asserted claims of the ’993 patent. As Omilia has previously raised with Nuance, Nuance still has no good faith basis to continue to assert these claims given these fundamental differences.

With regard to the doctrine of equivalents, Nuance’s infringement contentions fail to identify any aspect of any accused product that purportedly infringes any claim or any element of any asserted claim under the doctrine of equivalents. *See Intendis GMBH,* 822 F.3d at 1360 (“[e]ven when an accused product does not meet each and every claim element literally, it may nevertheless be found to infringe the claim ‘if there is ‘equivalence’ ***between the elements of the accused product or process and the claimed elements of the patented invention.***’”). Indeed, Nuance’s infringement contentions do not even identify what specific products allegedly infringe

under the doctrine of equivalents. Instead, Nuance's infringement contentions generically set forth boilerplate and include only conclusory allegations of infringement under the doctrine of equivalents. *See* Nuance's Response to Omilia's Interrogatory No. 9, Ex. B. This is insufficient as a matter of law. *See, e.g., Oil-Dri Corp. of Am.,* 2018 WL 1071443, *5 (striking doctrine of equivalents arguments where they "do not sufficiently address why the purported aspects of the Accused Products are equivalent and why any differences are insubstantial"). To the extent Nuance is permitted to supplement its infringement contentions with a doctrine of equivalents infringement theory, Omilia reserves the right to respond.

Omilia's investigation is ongoing and Omilia reserves the right to supplement or amend its responses based on additional discovery, or changes or supplements to Nuance's infringement allegations.

V. SECOND SUPPLEMENTAL INVALIDITY CONTENTIONS

In addition to its prior Invalidity Contentions, which Omilia NLS hereby reaffirms and reincorporates herein, Omilia NLS further states as follows:

On August 24, 2020, Nuance narrowed the asserted claims to claims to 1, 14, and 27 of the '925 patent and claims 17 and 19 of the '993 patent. Pursuant to the Court's scheduling order, Omilia provides further detail on its bases for invalidity of the remaining asserted claims. Dkt. No. 101. These supplemental contentions are focused on only these five claims. To the extent Nuance's asserted claims change, Omilia reserves the right to further supplement its contentions in response. Omilia NLS incorporates its prior invalidity disclosures (January 17, 2020 and June 9, 2020) and further explains the basis for invalidity below.

These invalidity arguments, including the prior art references and their use described below are merely representative and this list is not meant to limit Omilia NLS' invalidity arguments in this case. Omilia NLS reserves the right to use any of the described references to further support

the arguments identified. For example, Omilia NLS may use any of the current or previously disclosed references to describe the state of the art, meaning of certain terms, as well as elements that were conventional and routine during the relevant timeframe. Omilia NLS reserves the right to supplement these contentions as the case proceeds as well as in connection with expert discovery, in particular to provide additional details, if new prior art is discovered through discovery, or as Nuance further describes its infringement theories or interpretation of the claims or responds to Omilia NLS' invalidity allegations.

A. The '925 Patent

As described in Exhibit A (now broken into Exhibits A-1 to A-5), at least the prior art references below render the claims of the '925 Patent invalid.

- United States Patent No. 6,912,499, Sabourin et al., filed August 31, 1999 ("Sabourin").
- United States Patent No. 6,336,108, Thiesson et al., filed December 23, 1998 ("Thiesson").
- United States Patent No. 7,216,079, Barnard et al., filed November 2, 1999 ("Barnard").
- United States Patent Publication No. 2008-0147404, Liu et al., priority to May 15, 2000 ("Liu").
- United States Patent No. 6,151,574, Lee et al., filed September 8, 1998 ("Lee").
- Duchateau et al., *A Novel Node Splitting Criterion in Decision Tree Construction for Semi-Continuous HMMS*, 5th European Conference on Speech Communication and Technology EUROSPEECH '97 (1997) ("Duchateau").
- Schultz, et al., *Japanese LVCSR on the Spontaneous Scheduling Task with JANUS-3*, Eurospeech, Rhodes 1997 ("Schultz").
- M. Finke, et al., *Wide Context Acoustic Modeling in Read vs. Spontaneous Speech*, Proc. Of ICASSP, Munich 1997 ("Finke").
- V. Fischer, et al., *Speaker-Independent Upfront Dialect Adaptation in a Large Vocabulary Continuous Speech Recognizer*, 5th International Conference on Spoken Language Processing, Sydney, Australia 1998 ("Fischer").
- United States Patent No. 6,324,510, Waibel et al., filed November 6, 1998 ("Waibel").
- United States Patent No. 6,789,061, Fischer et al., filed August 14, 2000 (the "'061 Patent").
- R. Singh, et al., *Domain Adduced State Tying For Cross-Domain Acoustic Modeling*, Proc. Of the 6th Europ. Conf. on Speech Communication and Technology, Budapest (1999).

Claims 1, 14 and 25 of the ’925 patent are anticipated under § 102 or rendered obvious under § 103 in light at least the following bases:

1. **Thiesson** in combination with Sabourin or Schultz renders claims 1 and 14 of the ’925 patent obvious. In addition, Thiesson in combination with Sabourin renders claim 27 of the ’925 patent obvious. Thiesson describes a computerized method of “bootstrapping” or adapting a first speech recognizer to a second speech recognizer using specialized training data. It would have been obvious to combine Thiesson’s disclosure with Sabourin or Schultz, which disclose that the adaptation of a first speech recognizer to a second speech recognizer modifies the decision network and phonetic contexts of the first recognizer to create a second recognizer using specialized training data. It further would have been obvious to combine Sabourin disclosure concerning the resulting multilingual recognizer using a recognizer and training data from different languages. **Exhibit A-1** details the exemplary disclosures of Thiesson and combination references on an element-by-element basis and describes the specific motivations to combine Thiesson with those references.

2. **Sabourin** anticipates claims 1, 14 and 27 of the ’925 patent. In addition, Sabourin alone or in combination with Singh renders claims 1 and 14 of the ’925 patent obvious. Sabourin describes a computerized method of adapting a first language “speech model” to create a multilingual speech model by adjusting the phonetic contexts of the first model using training data from a second language. While the use of a decision tree is disclosed in Sabourin, the use of a decision tree in speech recognition would also have been obvious based on Sabourin’s disclosure alone or in combination with Singh, which describes the established and known method of using decision trees in speech recognizers. **Exhibit A-2** details the exemplary disclosures of Sabourin and combination references on an element-by-element basis and describes the specific motivations to combine Sabourin with those references.

3. **Schultz** anticipates or renders obvious claims 1, 14 and 27 of the ’925 patent. In addition, Schultz alone or in combination with Sabourin and Waibel render claims 1, 14 and 27 of the ’925 patent obvious. Schultz describes a computerized method of adapting a German speech recognizer to a Japanese speech recognizer by adjusting the phonetic contexts of the German recognizer using Japanese training data. It would have been obvious based on Schultz’s disclosure alone or in combination with Sabourin’s disclosure to create a multilingual recognizer. Sabourin describes a similar process of adapting phonetic contexts of a first recognizer with training data from a second language as described in Schultz. In addition, while Schultz discloses a computerized method for its adaptation, it also would have been obvious to combine Schultz’s disclosure with Sabourin and Waibel, which all describe using computers to automate the adaptation process. **Exhibit A-3** details the exemplary disclosures of Schultz and combination references on an element-by-element basis and describes the specific motivations

to combine Schultz with those references.

4. **Waibel** alone or in combination with Sabourin or Schultz render claim 27 of the ’925 patent obvious. Waibel describes a computerized method of adapting the decision network of a first speech recognizer to create a second speech recognizer by adjusting the phonetic contexts of the first decision network based on domain specific training data. It also would have been obvious to combine Waibel with Sabourin’s disclosure concerning the resulting multilingual recognizer, which both describe a similar process of adapting phonetic contexts of a first recognizer with second language training data. **Exhibit A-4** details the exemplary disclosures of Waibel and combination references on an element-by-element basis and describes the specific motivations to combine Waibel with those references.

5. **Fischer (’061 patent).** Claims 6 or Claim 15 of the ’061 patent render claims 1, 14, and 27 of the ’925 patent invalid for obvious type double patenting. Claims 6 or Claim 15 of the ’061 patent recite the same or an obvious variant of what is recited in claims 1, 14 and 27 of the ’925 patent. These claims are obvious in light of claims 6 or 15 of the ’061 patent alone or in combination with Sabourin or Schultz. The ’061 patent claims describes a computerize method of adapting a first recognizer to a second recognizer using domain specific training data. A resulting multilingual recognizer is an obvious variant. Moreover, it would have been obvious to combine these claims with Sabourin, which discloses a resulting multilingual recognizer using a first speech recognizer and training data from different languages. **Exhibit A-5** details the exemplary disclosures of ’061 patent and combination references on an element-by-element basis and describes the specific motivations to combine the ’061 patent with those references.

Omilia’s prior disclosures concerning the ’925 patent invalidity bases under 35 U.S.C. § 101 and 112 still apply. In addition, based on the court’s claim construction, the ’925 Patent fails to describe or enable what it means to be “multilingual” or “second language” as asserted by Nuance. While Omilia believes Nuance’s interpretations contradict the Court’s *Markman* order, these interpretations also render claim 27 invalid under 35 U.S.C. § 112.

B. The ’993 Patent

As described in Exhibit B (now broken into Exhibits B-1 to B-3), at least the prior art references below render the claims of the ’993 Patent invalid.

- Mirjam Wester, *Pronunciation Variation Modeling for Dutch Automatic Speech Recognition*, 2002 (“Wester”) and exemplary publications incorporated and discussed therein:

  - Kessens et al., *Improving the performance of a Dutch CSR by modeling within-word and cross-word pronunciation variation*, Speech Communication 29 (1999) 193-07 (incorporated as pp. 49-65 in Wester) ("Kessens")
  - Wester, *Pronunciation Modeling for ASR – Knowledge-based and Data-derived Methods*, 2001 (incorporated as pp. 97-122 in Wester) ("Wester2")
- Steinbiss et al., *The Philips research system for large-vocabulary continuous-speech recognition*, Proc. of 3rd European Conference on Speech Communication and Technology EUROSPEECH '93, 2125-28 (1993) ("Steinbiss")
- Kessens et al., *Improving the performance of a Dutch CSR by modeling within-word and cross-word pronunciation variation*, Speech Communication 29 (1999) 193-07 ("Kessens")
- Jain, et al., *Creating Speaker-Specific Phonetic Templates with a Speaker-Independent Phonetic Recognizer: Implications for Voice Dialing*, IEEE International Conference on Acoustics, Speech, and Signal Processing Conference Proceedings, 881-884 (1996) ("Jain")
- Helmer Strik, *Modeling pronunciation variation for ASR: A survey of the literature*, 1999 ("Strik")
- Florian Schiel, et al., *Statistical Modelling of Pronunciation: It's not the Model, It's the Data*, 1998 ("Schiel")

Claims 17 and 19 of the '993 patent are anticipated under § 102 or rendered obvious under § 103 in light at least the following bases:

1. **Wester** anticipates claims 17 and 19 of the '993 patent (Wester also incorporates disclosures from Wester2 and Kessens, which are companion articles relating to the same system). In addition, Wester alone or in combination with Jain and/or Steinbiss render claims 17 and 19 of the '993 patent obvious. Wester describes creating and using an improved speech recognition system with language models where pronunciations and their probabilities that are generated using phonetic transcribed information. Wester2 and Kessens similarly disclose creating and using an improved speech recognition system with language models where pronunciations and their probabilities are generated using phonetic transcribed information. Given this disclosure, it also would have been obvious to combine Wester's disclosure with Jain and/or Steinbiss. Wester discloses using phonetic transcribed speech that includes pronunciation variants. Jain describes using user-specific pronunciations to create phonetic transcribed speech that includes pronunciation variants. Wester is implemented using the computer-based system of, Steinbiss to create and use the modified language model. **Exhibit B-1** details the exemplary disclosures of Wester and combination references on an element-by-element basis and describes the specific motivations to combine Wester with those references.

2. **Schiel** anticipates claims 17 and 19 of the '993 patent. In addition, Schiel alone or

in combination with Jain and/or Steinbiss render claims 17 and 19 of the ’993 patent obvious. Schiel describes creating and using an improved speech recognition with pronunciation rules in the language models using pronunciations and their probabilities. These are created using digitized speech waves that are automatically segmented and converted into phonetic/phonemic symbols. Given this disclosure, it also would have been obvious to combine Schiel’s disclosure with Jain and/or Steinbiss. Schiel discloses using phonetic transcribed speech that includes pronunciation variants. Jain describes using user-specific pronunciations to create phonetic transcribed speech that includes pronunciation variants. Like Schiel, Steinbiss discloses the use of computers to create and use a speech recognition system. **Exhibit B-2** details the exemplary disclosures of Schiel and combination references on an element-by-element basis and describes the specific motivations to combine Schiel with those references.

3. **Strik** anticipates claims 17 and 19 of the ’993 patent. In addition, Strik alone or in combination with Jain, Steinbiss and/or Kessen render claims 17 and 19 of the ’993 patent obvious. Strik describes creating and using an improved speech recognition with language models with pronunciations and their probabilities. It also would have been obvious to combine Strik’s disclosure with Jain, Steinbiss and/or Kessen. Strik discloses using phonetic transcribed speech that includes pronunciation variants. Jain describes using user-specific pronunciations to create phonetic transcribed speech that includes pronunciation variants. Like Strik, Steinbiss discloses the use of computers to create and use a speech recognition system. Strik discloses that including many pronunciation variants may increase “confusability.” Kessens discloses using the most frequently occurring word sequences in order to protect against performance degradation. **Exhibit B-3** details the exemplary disclosures of Strik and combination references on an element-by-element basis and describes the specific motivations to combine Strik with those references.

Omilia’s prior disclosures concerning the ’993 patent invalidity bases under 35 U.S.C. § 101 and 112 for the still apply. In addition, based on the interpretation set forth by Nuance in its infringement contentions for the “approximating transcribed speech using a phonemic transcription dataset associated with a speaker to yield a language model, where the phonemic transcription dataset is based on a pronunciation model of the speaker,” the claims of the ’993 patent are further invalid under § 112 because it fails to provide sufficient disclosure or enablement for this limitation. Nuance’s contentions far exceed the scope of what is described and enabled in the ’993 patent.

Dated: October 30, 2020

Respectfully Submitted,

/s/ *Daniel S. Sternberg*

Kevin C. Adam (SBN 684955)
Daniel S. Sternberg (SBN 688842)
WHITE & CASE LLP
75 State Street, 24th Floor
Boston, MA 02109
(617) 979-9300
kevin.adam@whitecase.com
daniel.sternberg@whitecase.com

Of Counsel:
Dimitrios Drivas (*admitted pro hac vice*)
Raj Gandesha (*admitted pro hac vice*)
Stefan Mentzer (*admitted pro hac vice*)
John Padro (*admitted pro hac vice*)
WHITE & CASE LLP
1221 Avenue of the Americas
New York, NY 10020-1095
(212) 819-8286
ddrivas@whitecase.com
rgandesha@whitecase.com
smentzer@whitecase.com
john.padro@whitecase.com

Hallie Kiernan (*admitted pro hac vice*)
WHITE & CASE LLP
3000 El Camino Real
Two Palo Alto Square, Suite 900
Palo Alto, CA 94306
(650) 213-0300
hallie.kiernan@whitecase.com

Counsel for Omilia Natural Language Solutions, Ltd

CERTIFICATE OF SERVICE

I hereby certify that on October 30, 2020, the foregoing document was served by electronic mail to the following attorneys of record:

David J. Lender
Eric S. Hochstadt
WEIL, GOTSHAL & MANGES LLP
767 Fifth Avenue
New York, NY 10153
Telephone: 212-310-8000
David.lender@weil.com
Eric.hochstadt@weil.com

Jennifer Itzkoff
WOMBLE BOND DICKINSON (US) LLP
570 Atlantic Avenue, Suite 600
Boston, MA 02110
Telephone: 857-287-3142
Jennifer.itzkoff@wbd-us.com

Christian E. Mammen
Carrie Richey
WOMBLE BOND DICKINSON (US) LLP
1841 Page Mill Road, Suite 200
Palo Alto, CA 94304
Telephone: 408-341-3067
Chris.mammen@wbd-us.com
Telephone: 408-341-3060
Carrie.richey@wbd-us.com

Kristin Lamb
WOMBLE BOND DICKIINSON (US) LLP
811 Main Street, Suite 3130
Houston, TX 77002
Telephone: (346) 998-7843
Kristin.lamb@wbd-us.com

Christine H. Dupriest
WOMBLE BOND DICKINSON (US) LLP
271 – 17th Street, Suite 2400
Atlanta, GA 30363
Telephone: 404-962-7538
Christine.dupriest@wbd-us.com

/s/ Hallie Kiernan
Hallie Kiernan

APPENDIX 1

Entire Document Subject to Filing Under Seal

APPENDIX 2

Entire Document Subject to Filing Under Seal

APPENDIX A-1
Invalidity Claim Chart for U.S. Pat. No. 6,999,925 ('925 patent)
U.S. Pat. No. 6,336,108, Thiesson et al., filed December 23, 1998 ("Thiesson")[1]

On October 16, 2020, Nuance narrowed the asserted '925 patent claims to 1, 14 and 27. Nuance's Disclosure Limiting Asserted Claims of Phase 1 Patents. Claims 1, 14 and 27 of the '925 patent are anticipated and/or rendered obvious by Thiesson alone or in combination with at least the following references. The chart below discloses how prior art references identified by Omilia disclose, either expressly or inherently, and/or render obvious each of the asserted claims:

(1) U.S. Pat. No. 6,912,499, Sabourin et al., filed August 31, 1999 ("Sabourin")[2]

(2) Schultz, et al., *Japanese LVCSR on the Spontaneous Scheduling Task with JANUS-3*, Eurospeech, Rhodes 1997 ("Schultz")[3]

This invalidity claim chart is based on Omilia's present understanding of the asserted claims, the Court's August 6, 2020 Claim Construction Order on (ECF. No. 157), Nuance's initial and all subsequent supplemental Infringement Contentions, its July 7, 2020 Response to Omilia's Supplemental Non-Infringement and Invalidity Responses, Nuance's Response to Omilia's Interrogatory No. 9, and Omilia's investigation to date. Omilia is not adopting Nuance's apparent constructions, nor is Omilia admitting to the accuracy of any particular construction. Omilia reserves all rights to amend this invalidity claim chart if Nuance further amends its Infringement Contentions, or in response to any statements made by Nuance at any phase of this litigation. In addition, Discovery is still ongoing, and Omilia reserves its right to supplement its contentions with additional evidence as discovery continues.

The citations provided are exemplary and do not necessarily include each and every disclosure of the limitation in the references. Omilia has cited the most relevant portions of the identified prior art. Other portions of the identified prior art may additionally disclose, either expressly or inherently, and/or render obvious one or more limitations of the asserted claims. To provide context or aid in understanding the prior art and the state of the art at the time of the alleged invention, Omilia reserves the right to rely on: (1) uncited portions of the identified prior art; (2) other prior art not identified herein; (3) references that show the state of the art (irrespective of whether such references themselves qualify as prior art to the asserted patents); (4) factual testimony from the inventors or authors of the prior art references, or purveyors of prior art devices; and/or (5) expert testimony.

---

[1] Thiesson was filed on December 23, 1998 and is prior art at least under 35 U.S.C. § 102(a) & 102(e).

[2] Sabourin was filed on August 31, 1999 and constitutes prior art at least under 35 U.S.C. § 102(a) & 102(e).

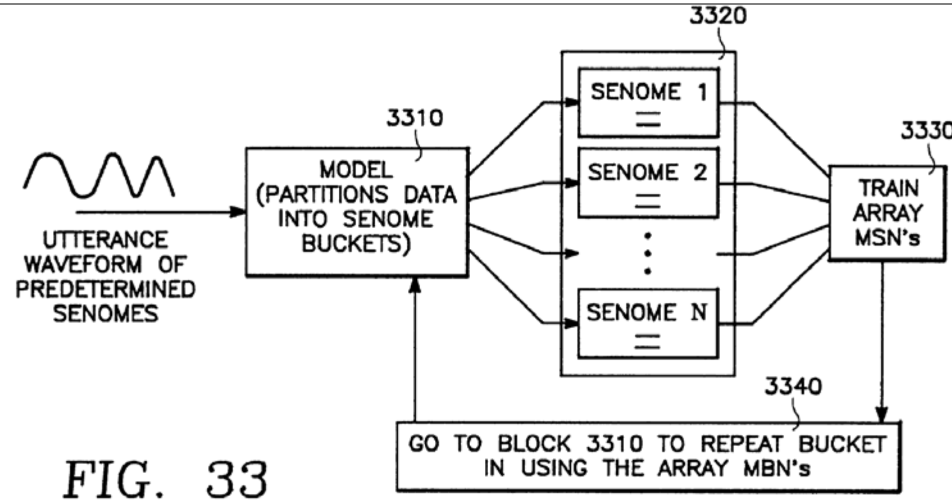
[3] Schultz was presented at Eurospeech 97 from September 22-25, 1997 and constitutes prior art at least under 35 U.S.C. § 102(a) & 102(b).

Citations to particular drawing or figure in the accompanying charts should be understood to include the description of the drawing or figure, as well as any text associated with the drawing or figure. Citations to particular text concerning a drawing or figure, should also be understood to encompasses that drawing or figure as well. The identified prior art expressly and/or inherently discloses the features of the asserted claims under all proposed constructions of the asserted claims. To establish the inherency of certain features of the prior art to invalidate the asserted claims, Omilia may rely on cited or uncited portions of the prior art, other documents, factual testimony, and expert testimony. To the extent that the Court finds that this reference does not explicitly teach certain limitations in the asserted claims, such limitations would have been inherent and/or obvious and Omilia reserves the right to supplement or amend this claim chart to address such a finding.

<u>’925 Patent</u>		
Claim 1		
1.pre.a	"A computerized method of automatically generating from a first speech recognizer a second speech recognizer"	Thiesson discloses a computerized method of automatically generating from a first speech recognizer a second speech recognizer. Thiesson discloses utilizing a personal computer system to run the invention. Thiesson describes using a fast bootstrap training procedure to adapt a first speech recognizer into a second speech recognizer. *See, e.g.*, "The advent of artificial intelligence within computer science has brought an abundance of decision-support systems. Decision-support systems are computer systems in which decisions, typically rendered by humans, are recommended and sometimes made. In creating decision-support systems, computer scientists seek to provide decisions with the greatest possible accuracy. Thus, computer scientists strive to create decision-support systems that are equivalent to or more accurate than a human expert. Applications of decision-support systems include medical diagnosis, troubleshooting computer networks, or other systems wherein a decision is based upon identifiable criteria." Thiesson at 1:18-29; *see* Thiesson claim 54. "With reference to FIG. 4, an exemplary system for implementing the invention includes a general purpose computing device in the form of a conventional personal computer 420, including a processing unit 421, a System memory 422, and a System buS 423 that couples various System components including the System memory to the processing unit 421. The system bus 423 may be any of several types of bus Structures including a memory bus or

		memory controller, a peripheral bus, and a local bus using any of a variety of bus architectures. The System memory includes read only memory (ROM) 424 and random access memory (RAM) 425. A basic input/output system 426 (BIOS), containing the basic process that helps to transfer information between elements within the personal computer 420, Such as during start-up, is stored in ROM 424. The personal computer 420 further includes a hard disk drive 427 for reading from and Writing to a hard disk, not shown, a magnetic disk drive 428 for reading from or writing to a removable magnetic disk 429, and an optical disk drive 430 for reading from or writing to a removable optical disk 431 Such as a CD ROM or other optical media. The hard disk drive 427, magnetic disk drive 428, and optical disk drive 430 are connected to the system bus 423 by a hard disk drive interface 432, a magnetic disk drive interface 433, and an optical drive interface 434, respectively. The drives and their associated computer-readable media provide nonvolatile Storage of computer readable instructions, data Structures, program modules and other data for the personal computer 420. Although the exemplary environment described herein employs a hard disk, a removable magnetic disk 429 and a removable optical disk 431, it should be appreciated by those skilled in the art that other types of computer readable media which can Store data that is accessible by a computer, Such as magnetic cassettes, flash memory cards, digital Video disks, Bernoulli cartridges, random access memories (RAMs), read only memories (ROM), and the like, may also be used in the exemplary operating environment. A number of program modules may be Stored on the hard disk, magnetic disk 429, optical disk 431, ROM 424 or RAM 425, including an operating system 435, one or more application programs 436, other program modules 437, and program data 438. A user may enter commands and information into the personal computer 420 through input devices such as a keyboard 440 and pointing device 442. Other input devices (not shown) may include a microphone, joystick, game pad, Satellite dish, Scanner, or the like. These and other input devices are often connected to the processing unit 421 through a Serial port interface 446 that is coupled to the System bus, but may be connected by other interfaces, Such as a parallel port, game port or a universal Serial bus (USB). A monitor 447 or other type of display device is also connected to the System bus 423 via an interface, Such as a video adapter 448. In addition to the monitor, personal computers typically include other peripheral output devices (not shown), Such as speakers and printers.”

		Thiesson at 10:50-11:36. "Preferably, there are 20 HSBNs 3050 in each MBN 3020 (i.e., m=20) so that the hidden class variable C has 20 states. The hidden class variable C accounts for other variables not included in the model, Such as, for example, different speech dialects among different clusters of Speakers." Thiesson at 37:63-67. "FIG. 33 illustrates a bootstrap training procedure involving the entire speech recognition system of FIG. 30 including the language model 3040. In this case, the training data consists of an input utterance waveform representing a predetermined (a priori known) sequence of Senones. The speech recognition system of FIG. 30 partitions the data (block 3310 of FIG. 33) by assigning each successive acoustic observation vector to a particular Senone "bucket' (block 3320 of FIG. 33). Initially, when this step is performed the speech recognition system of FIG. 30 can employ a conventional Structure as an acoustic model in place of the acoustic model 3010 consisting of the array of MBNs of the present invention. Once the 33-dimensional acoustic observation vectors have been assigned to various Senone buckets by the step of block 3320, the resulting data is used to train the array of MBNs 3010 of FIG. 30 using the training procedures described in this specification (block 3330 of FIG. 33). The array of MBNs 3010 is then used as the acoustic model in place of the conventional acoustic model in the speech recognition system of FIG. 30, and the data partitioning step of block 3310 is then repeated (block 3340 of FIG. 33). This may result in a change in the way in which Some of the acoustic observation vectors are bucketed, resulting in a more correct Set of training data. Then, using this improved training data Set, the array of MBNS 3010 is re-trained using the training procedures described in this specification. This latter Step is a repetition of the step of block 3330 of FIG.33. Thereafter, as more new acoustic observations are received, the bucketing Step of block 3310 is carried out for the incoming acoustic observation vectors, and the entire process of FIG.33 recycles in this manner." Thiesson at 38:53-39:16; *see also* Fig 33.

		 FIG. 33
1.pre.b	"said first speech recognizer comprising a first acoustic model with a first decisions network and corresponding first phonetic contexts"	Thiesson discloses a first speech recognizer with a first acoustic model, first decision network and corresponding phonetic contexts. Thiesson describes using conventional structures in its fast bootstrapping procedure, and also the use of Bayesian networks of acoustic observations, which make up an acoustic model. *See, e.g.,* "The invention performs speech recognition using an array of mixtures of Bayesian networks. A mixture of Bayesian networks (MBN) consists of plural hypothesis-specific Bayesian networks (HSBNs) having possibly hidden and observed variables. A common external hidden variable is associated with the MBN, but is not included in any of the HSBNs. The number of HSBNs in the MBN corresponds to the number of states of the common external hidden variable, and each HSBN models the world under the hypothesis that the common external hidden variables is in a corresponding one of those states. In accordance with the invention, the MBNs encode the probabilities of observing the sets of acoustic observations given the utterance of a respective one of said parts of speech. Each of the HSBNs encodes the probabilities of observing the sets of acoustic observations given the utterance of a respective one of the parts of speech and given the hidden common variable being in a particular state. Each HSBN has nodes corresponding to the elements of the acoustic observations. These nodes store probability parameters corresponding to the probabilities with causal links representing dependencies between ones of said node."

		Thiesson at Abstract. "FIG. 33 illustrates a bootstrap training procedure involving the entire speech recognition system of FIG. 30 including the language model 3040. In this case, the training data consists of an input utterance waveform representing a predetermined (a priori known) sequence of Senones. The speech recognition system of FIG. 30 partitions the data (block 3310 of FIG. 33) by assigning each successive acoustic observation vector to a particular Senone "bucket' (block 3320 of FIG. 33). Initially, when this step is performed the speech recognition system of FIG. 30 can employ a conventional Structure as an acoustic model in place of the acoustic model 3010 consisting of the array of MBNs of the present invention. Once the 33-dimensional acoustic observation vectors have been assigned to various Senone buckets by the step of block 3320, the resulting data is used to train the array of MBNs 3010 of FIG. 30 using the training procedures described in this specification (block 3330 of FIG. 33). The array of MBNs 3010 is then used as the acoustic model in place of the conventional acoustic model in the speech recognition system of FIG. 30, and the data partitioning step of block 3310 is then repeated (block 3340 of FIG. 33). This may result in a change in the way in which Some of the acoustic observation vectors are bucketed, resulting in a more correct Set of training data. Then, using this improved training data Set, the array of MBNS 3010 is re-trained using the training procedures described in this specification. This latter Step is a repetition of the step of block 3330 of FIG.33. Thereafter, as more new acoustic observations are received, the bucketing Step of block 3310 is carried out for the incoming acoustic observation vectors, and the entire process of FIG.33 recycles in this manner." Thiesson at 38:53-39:16; *see also* Fig 33.

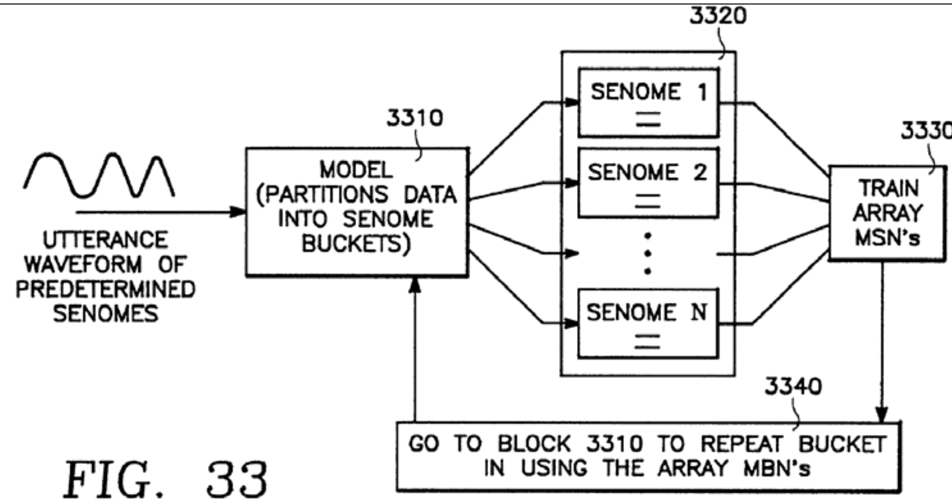
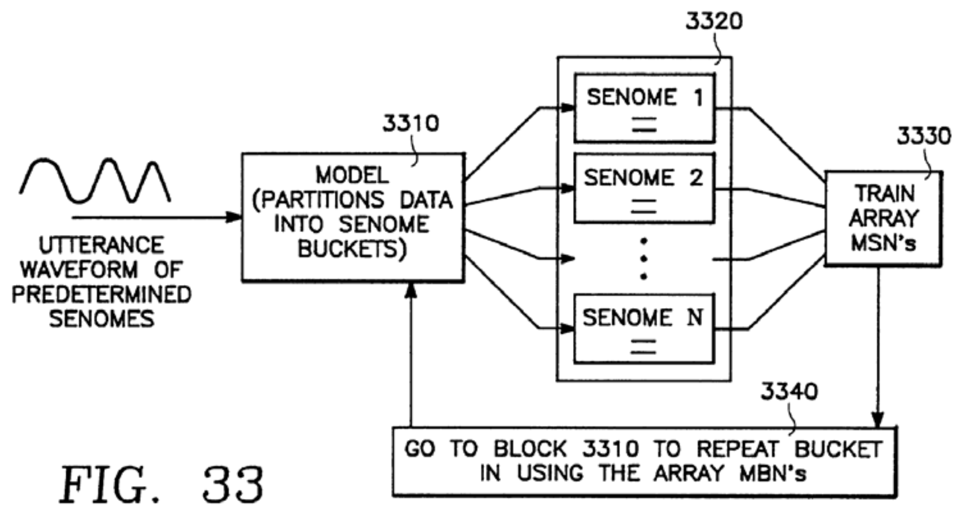


FIG. 33

In the alternative, Thiesson suggests that the acoustic model may convey a conventional structure and that one such conventional structure is a decision tree. A decision tree is a decision network and includes by virtue of its structure, phonetic context for each leaf of the network. *See, e.g.*,

"A tree data Structure is an acyclic, undirected graph where each vertex is connected to each other vertex via a single path. The graph is acyclic in that there is no path that both emanates from a vertex and returns to the same vertex, where each edge in the path is traversed only once. FIG. 3C depicts an example tree data structure 330 that stores into its leaf vertices 336-342 the probabilities shown in table 320 of FIG. 3B. Assuming that a decision-support system performs probabilistic inference with X's value being 0 and Y S value being 1, the following Steps occur to access the appropriate probability in the tree data Structure 330: First, the root vertex 332, vertex X, is accessed, and its value determines the edge or branch to be traversed. In this example, X's value is 0, so edge 344 is traversed to vertex 334, which is vertex Y. Second, after reaching vertex Y, the value for this vertex determines which edge is traversed to the next vertex. In this example, the value for vertex Y is 1, so edge 346 is traversed to vertex 338, which is a leaf vertex. Finally, after reaching the leaf vertex 338, which stores the probability for Z equaling 0 when X=0 and Y =1, the appropriate probability

		can be accessed. AS compared to a table, a tree is a more efficient way of Storing probabilities in a node of a Bayesian network, because it requires less Space. However, tree data Structures are inflexible in the Sense that they can not adequately represent relationships between probabilities. For example, because of the acyclic nature of tree data Structures, a tree cannot be used to indicate Some types of equality relationships where multiple combinations of the values of the parent Vertices have the same probability (i.e., refer to the same leaf vertex). This inflexibility requires that multiple vertices must Sometimes store the same probabilities, which is wasteful. It is thus desirable to improve Bayesian networks with tree distributions." Thiesson at 5:6-39; *see also* Thiesson Fig 3C.
1.pre.c	"and said second speech recognizer being adapted to a specific domain, said method comprising:"	Thiesson discloses a second speech recognizer being adapted by training data. *See also claim 1.pre.b. See also*, "FIG. 33 illustrates a bootstrap training procedure involving the entire speech recognition system of FIG. 30 including the language model 3040. In this case, the training data consists of an input utterance waveform representing a predetermined (a priori known) sequence of senones. The speech recognition system of FIG. 30 partitions the data (block 3310 of FIG. 33) by assigning each successive acoustic observation vector to a particular senone "bucket' (block 3320 of FIG. 33). Initially, when this step is performed the speech recognition system of FIG. 30 can employ a conventional structure as an acoustic model in place of the acoustic model 3010 consisting of the array of MBNs of the present invention. Once the 33-dimensional acoustic observation vectors have been assigned to various senone buckets by the step of block 3320, the resulting data is used to train the array of MBNs 3010 of FIG. 30 using the training procedures described in this specification (block 3330 of FIG. 33). The array of MBNs 3010 is then used as the acoustic model in place of the conventional acoustic model in the speech recognition system of FIG. 30, and the data partitioning step of block 3310 is then repeated (block 3340 of FIG. 33). This may result in a change in the way in which Some of the acoustic observation vectors are bucketed, resulting in a more correct Set of training data. Then, using this improved training data Set, the array of MBNS 3010 is re-trained using the training procedures described in this specification. This latter Step is a repetition of the step of block 3330 of FIG.33. Thereafter, as more new acoustic observations

		are received, the bucketing Step of block 3310 is carried out for the incoming acoustic observation vectors, and the entire process of FIG.33 recycles in this manner." Thiesson at 38:53-39:16; *see also* Thiesson Fig. 33.  FIG. 33 In addition, a POSITA would have recognized that at the time of the invention, training a speech recognizer to a specific domain was well known in the art and used in a variety of situations. A POSITA would have looked to other references to describe domain specific training data, such as Sabourin and Schultz. *See* Schultz at Abstract, Section 3.3; Sabourin at Abstract, 2:23-28, 6:45-7:8. A POSITA would have been motivated to combine Thiesson with any one of these references. Moreover, given this disclosure, the use of domain specific training data was obvious in light of Thiesson alone, or in combination with any of the above references. The motivation to combine these references would at least include: • Combining prior art elements according to known methods to yield predictable results (use of domain specific training data and adaptation based on that training data was known);

		• Simple substitution of one known element for another to obtain predictable results (substituting just the training data); • Use of known technique to improve similar devices (methods, or products) in the same way (same technique of adapting the recognizer to new domain training data)); • Applying a known technique to a known device (method, or product) ready for improvement to yield predictable results (adapted recognizers were known using similar known techniques for predictable results); • "Obvious to try"—choosing from a finite number of identified, predictable solutions, with a reasonable expectation of success (obvious to try to use new training data to a specific domain); • Market forces and benefits associated with the known benefits of multilingual recognizers were predictable to POSA • Teaching of prior art would have lead a POSA to combine the references to arrive at a multilingual recognizer. It would be obvious to a person having ordinary skill in the art to combine Thiesson with any of the above references to disclose utilizing domain specific training data to adapt the acoustic model. All of the references are in the same field of art. A person having ordinary skill in the art would be motivated to combine Thiesson with any of the references. A person having ordinary skill in the art considering Thiesson's disclosure that the hidden variables may be dialects would be motivated to seek out a system that adapts the first decision network using domain specific training data to better account for things like the hidden dialects, as disclosed by the above references. A person having ordinary skill in the art would have a reasonable expectation of success in implementing the teaching of Thiesson to adapt the acoustic model based on domain specific training data as provided by the above references because the combination involves the predictable use of prior art elements according to their established functions.
1.a.1	"based on said first acoustic model, generating a second acoustic model with a	Thiesson in combination with the references below renders this limitation obvious and discloses that based on said first acoustic model, generating a second acoustic model with a second decision network and corresponding second phonetic contexts for said second speech

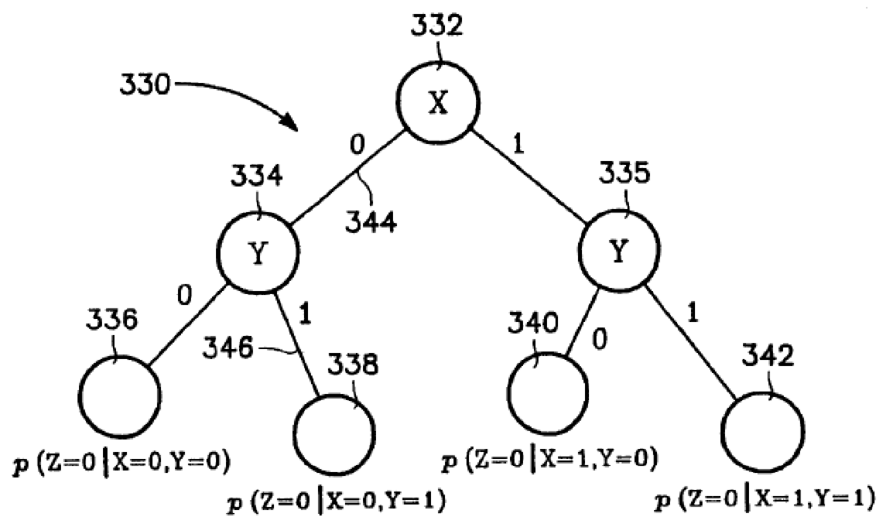
	second decision network and corresponding second phonetic contexts for said second speech recognizer by re-estimating said first decision network and said corresponding first phonetic contexts based on domain-specific training data,"	recognizer by re-estimating said first decision network and said corresponding first phonetic contexts based on domain-specific training data. *See, e.g.*, "FIG. 33 illustrates a bootstrap training procedure involving the entire speech recognition system of FIG. 30 including the language model 3040. In this case, the training data consists of an input utterance waveform representing a predetermined (a priori known) sequence of senones. The speech recognition system of FIG. 30 partitions the data (block 3310 of FIG. 33) by assigning each successive acoustic observation vector to a particular senone "bucket' (block 3320 of FIG. 33). Initially, when this step is performed the speech recognition system of FIG. 30 can employ a conventional Structure as an acoustic model in place of the acoustic model 3010 consisting of the array of MBNs of the present invention. Once the 33-dimensional acoustic observation vectors have been assigned to various senone buckets by the step of block 3320, the resulting data is used to train the array of MBNs 3010 of FIG. 30 using the training procedures described in this specification (block 3330 of FIG. 33). The array of MBNs 3010 is then used as the acoustic model in place of the conventional acoustic model in the speech recognition system of FIG. 30, and the data partitioning step of block 3310 is then repeated (block 3340 of FIG. 33). This may result in a change in the way in which Some of the acoustic observation vectors are bucketed, resulting in a more correct Set of training data. **Then, using this improved training data set, the array of MBNs 3010 is re-trained using the training procedures described in this specification.** This latter step is a repetition of the step of block 3330 of FIG.33. Thereafter, as more new acoustic observations are received, the bucketing Step of block 3310 is carried out for the incoming acoustic observation vectors, and the entire process of FIG.33 recycles in this manner." Thiesson at 38:53-39:16 (emphasis added). "However, if an alternative embodiment of the present invention is used where cycles are allowed to be introduced into the Bayesian network, then complete splits are performed on all nodes in the Bayesian network other than the parent of the leaf node. When performing a complete Split, the Bayesian network generator Selects one of the non-descendent nodes described above and replaces the leaf node in the decision graph with a vertex that corresponds to the Selected non-descendent node. Then, new leaves are created which depend from the newly created vertex; one leaf vertex is created for each value of the newly added vertex. For example, if the leaf vertex 1908 of the decision graph 1904 of FIG. 27A

		had a complete split performed on the age node, the resulting decision graph appears in FIG. 27B where the leaf 1908 of FIG. 27A is replaced with age vertex 1918 of FIG. 27B and leaves 1920-1926 are created, one for each value of the age vertex (i.e., each State of the age node of the Bayesian network). Each complete split on a particular non-descendent node generates a new decision graph which is Stored. To conserve Space, an exemplary embodiment Stores an identification of the change and not the entire decision graph. After performing a complete split, the Bayesian network generator performs a binary split if the number of States is greater than two (step 1844). In this step, a binary split is performed on the leaf for all nodes that are not descendants of the identified node as reflected in the Bayesian network and for all values for these non-descendent nodes. AS Stated above, this restriction is enforced to prevent the addition of cycles into the Bayesian network. However, an alternative embodiment does not enforce this restriction. In a binary Split operation, a leaf is replaced with a vertex that corresponds to one of the non-descendant nodes, and two leaves are generated from the newly created vertex node: one of the leaves contains a Single value and the other leaf contains all other values. For example, in the decision graph 1904 of FIG. 27A, if leaf 1908 had a binary split performed on the age variable, the leaf 108 of FIG. 27A would be replaced with age vertex 1930 as shown in FIG. 27C and two leaves 1932 and 1934 would be generated for that vertex. The first leaf 932 would contain one value (e.g., 1) and the Second leaf 1934 would be for all other values of the age vertex 1930 (e.g., 0, 2 and 3). As stated above, the binary splits on the leaf will be performed for all non-descendent nodes and for each value of each non-descendent node. Thus, when a node has n values, a binary Split is performed on this node n times. For example, Since the age node has four values, four splits would occur: (1) one leaf would have a value of 0, and the other leaf would have a value of 1, 2, or 3; (2) one leaf would have a value of 1, and the other leaf would have a value of 0, 2, or 3; (3) one leaf would have a value of 2, and the other leaf would have a value of 0, 1, or 3; (4) one leaf would have a value of 3, and the other leaf would have a value of 0, 1, or 2. The Bayesian network generator Stores identifications of the changes reflected by these binary splits. After performing a binary split, the Bayesian network generator merges all pairs of leaf nodes together (step 1846). In this step, the Bayesian network generator generates a number of new decision graphs by merging the leaf node selected in step 1840 with each other leaf

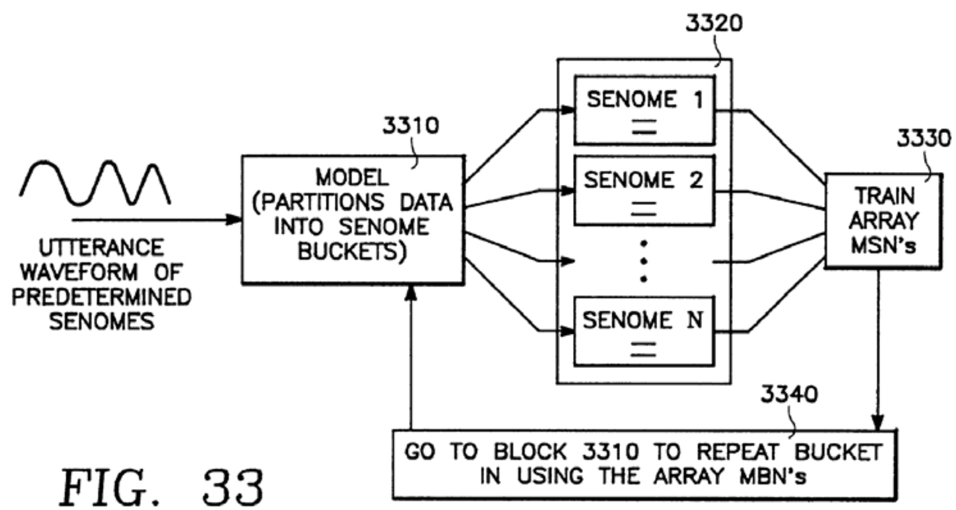
		node to form a Single vertex. For example, with respect to the decision graph 1904 of FIG. 27A, leaf 1908 and leaf 1912 can be merged into a single leaf 1938 as depicted in FIG.27D. After merging all pairs of leaf nodes, the Bayesian network generator determines if the decision graph has more leaves for processing. If So, processing continues to Step 1840. Otherwise, processing ends. Although the exemplary embodiment is described as performing a complete split, a binary split, and a merge, one skilled in the art will appreciate that other operations can be performed." Thiesson at 33:59-34:61; *see also* Thiesson at Figs. 27A, 27B, 27C, and 27D. It would have been obvious to combine Thiesson's adaptation of a first recognizer with at least Sabourin and Schultz, which describe the same type of adaptation where the phonetic contexts and decision network structure is modified. A POSITA would have looked to similar art in the field and recognized it was well known to add or delete nodes and adapt the phonetic contexts in the decision network based on at least Sabourin and Schultz. Moreover, given this disclosure, the re-estimation of the first speech recognizer was obvious in light of Thiesson in combination with Sabourin or Schultz. The motivation to combine these references would at least include: • Combining prior art elements according to known methods to yield predictable results (adding nodes to account for new phones in a domain or deleting unused phones were known using similar techniques for predictable results); • Simple substitution of one known element for another to obtain predictable results; • Use of known technique to improve similar devices (methods, or products) in the same way (same technique of adapting the decision network when expanding the domain of the recognizer to include new phones or deleting unused phones was known)); • Applying a known technique to a known device (method, or product) ready for improvement to yield predictable results (adding and deleting nodes in the decision

		network during adaptation were known using similar known techniques for predictable results); • "Obvious to try"—choosing from a finite number of identified, predictable solutions, with a reasonable expectation of success (obvious to add nodes for new sounds or remove nodes for unused phones); • Market forces and benefits associated with the known benefits of adding and deleting nodes were predictable to POSA • Teaching of prior art would have lead a POSA to combine the references to arrive at an adaptation of a recognizer that includes adding or deleting nodes. For example, Sabourin and Schultz explicitly perform re-estimation by deleting unused nodes, adding nodes, pruning nodes and merging nodes in the decision network. It would be obvious to a person having ordinary skill in the art to combine Thiesson with Sabourin or Schultz to disclose deleting nodes, adding nodes, pruning nodes and merging nodes in the decision network. Thiesson, Sabourin, and Schultz are in the same field of art. A person having ordinary skill in the art would be motivated to combine Thiesson with Sabourin or Schultz. A person having ordinary skill in the art considering Thiesson's disclosure that "the array of MBNs 3010 is re-trained" would be motivated to seek out other systems that train the first decision network, including by adding and deleting nodes, as disclosed by Sabourin and Schultz, to attempt to achieve an even further improved system. A person having ordinary skill in the art would have a reasonable expectation of success in implementing the teaching of Thiesson to re-estimate the first decision network through addition of nods and deletion of nodes as provided by Sabourin and Schultz because the combination involves the predictable use of prior art elements according to their established functions. *See* Sabourin at Abstract, 2:57-3:6, 3:52-4:7, 4:16-45, 5:66-6:44, 6:45-7:8, 8:49-10:23; Schultz at Abstract, Section 3.3., Section 3.4.
1.a.2	"wherein said first decision network and said second decision network utilize a	Thiesson discloses that the first and second decision networks may utilize phonetic decision trees to perform speech operations.

	phonetic decision free [sic] to perform speech recognition operations”	“FIG. 33 illustrates a bootstrap training procedure involving the entire speech recognition system of FIG. 30 including the language model 3040. In this case, the training data consists of an input utterance waveform representing a predetermined (a priori known) sequence of senones. The speech recognition system of FIG. 30 partitions the data (block 3310 of FIG. 33) by assigning each successive acoustic observation vector to a particular senone “bucket' (block 3320 of FIG. 33). Initially, when this step is performed the speech recognition system of FIG. 30 **can employ a conventional Structure** as an acoustic model in place of the acoustic model 3010 consisting of the array of MBNs of the present invention. Once the 33-dimensional acoustic observation vectors have been assigned to various senone buckets by the step of block 3320, the resulting data is used to train the array of MBNs 3010 of FIG. 30 using the training procedures described in this specification (block 3330 of FIG. 33). The array of MBNs 3010 is then used as the acoustic model in place of the conventional acoustic model in the speech recognition system of FIG. 30, and the data partitioning step of block 3310 is then repeated (block 3340 of FIG. 33). This may result in a change in the way in which Some of the acoustic observation vectors are bucketed, resulting in a more correct Set of training data. Then, using this improved training data set, the array of MBNs 3010 is re-trained using the training procedures described in this specification. This latter step is a repetition of the step of block 3330 of FIG.33. Thereafter, as more new acoustic observations are received, the bucketing Step of block 3310 is carried out for the incoming acoustic observation vectors, and the entire process of FIG.33 recycles in this manner.” Thiesson at 38:53-39:16 (emphasis added). Thiesson discloses using conventional structures in the fast bootstrapping procedure and describes a decision tree as a conventional structure. Thus, Thiesson teaches at least that a decision tree can be used in its fast bootstrapping adaptation procedure. *See, e.g.,* “A tree data Structure is an acyclic, undirected graph where each vertex is connected to each other vertex via a single path. The graph is acyclic in that there is no path that both emanates from a vertex and returns to the same vertex, where each edge in the path is traversed only once. FIG. 3C depicts an example tree data structure 330 that stores into its leaf vertices 336-342 the probabilities shown in table 320 of FIG. 3B. Assuming that a decision-support system performs probabilistic inference with X's value being 0 and Y S value being 1, the following Steps occur to access the appropriate probability in the tree data Structure 330: First, the root vertex 332, vertex X, is accessed, and its value determines the edge or branch

		to be traversed. In this example, X's value is 0, so edge 344 is traversed to vertex 334, which is vertex Y. Second, after reaching vertex Y, the value for this vertex determines which edge is traversed to the next vertex. In this example, the value for vertex Y is 1, so edge 346 is traversed to vertex 338, which is a leaf vertex. Finally, after reaching the leaf vertex 338, which stores the probability for Z equaling 0 when X=0 and Y =1, the appropriate probability can be accessed. AS compared to a table, a tree is a more efficient way of Storing probabilities in a node of a Bayesian network, because it requires less Space. However, tree data Structures are inflexible in the Sense that they can not adequately represent relationships between probabilities. For example, because of the acyclic nature of tree data Structures, a tree cannot be used to indicate Some types of equality relationships where multiple combinations of the values of the parent Vertices have the same probability (i.e., refer to the same leaf vertex). This inflexibility requires that multiple vertices must Sometimes store the same probabilities, which is wasteful. It is thus desirable to improve Bayesian networks with tree distributions." Thiesson at 5:6-39; *see also* Fig. 3C. 

		"FIG. 3C depicts a tree data structure containing the probabilities for one of the nodes of the Bayesian network of FIG 3A." Thiesson at 8:61-63. "In the exemplary embodiment, a Bayesian network (i.e., the initial network or the test network 608) is stored in memory as a tree data Structure where each node in the tree data Structure corresponds to a node in the Bayesian network. The arcs of the Bayesian network are implemented as pointers from one node in the tree data Structure to another node. In addition, the probabilities for each node in the Bayesian network are Stored in the corresponding node in the tree data Structure." Thiesson at 22:17-25.
1.b.1	"wherein the number of nodes in the second decision network is not fixed by the number of nodes in the first decision network,"	Thiesson discloses a process of modifying the decision network to split and merge nodes. Combining this adaptation process with the fast bootstrap in FIG 33, would be obvious because Thiesson explicitly contemplates that the training procedure of the patent will be applied after the partitioning step in the fast bootstrap procedure. Thiesson's disclosure of splitting and merging nodes alone or further combined with references that add or delete nodes, for example Sabourin and Schultz (*see* claim 1.a.1) necessarily means that the number of nodes in the second decision network is not fixed by the number of nodes in the first decision network. Thus, Thiesson in combination with the references as described in claim 1.a.1 discloses that "the number of nodes in the second decision network is not fixed by the number of nodes of the first decision network."
1.b.2	"and wherein said re-estimating comprises partitioning said training data using said first decision network of said first speech recognizer."	Thiesson discloses that the re-estimation comprises partitioning said training data using said first decision network of said first speech recognizer. Thiesson describes the speech recognition unit partitioning the training data. *See, e.g.*, "FIG. 33 illustrates a bootstrap training procedure involving the entire speech recognition system of FIG. 30 including the language model 3040. In this case, the training data consists of an input utterance waveform representing a predetermined (a priori known) sequence of

		senones. The speech recognition system of FIG. 30 partitions the data (block 3310 of FIG. 33) by assigning each successive acoustic observation vector to a particular senone "bucket' (block 3320 of FIG. 33). Initially, when this step is performed the speech recognition system of FIG. 30 can employ a conventional Structure as an acoustic model in place of the acoustic model 3010 consisting of the array of MBNs of the present invention. Once the 33-dimensional acoustic observation vectors have been assigned to various senone buckets by the step of block 3320, the resulting data is used to train the array of MBNs 3010 of FIG. 30 using the training procedures described in this specification (block 3330 of FIG. 33). The array of MBNs 3010 is then used as the acoustic model in place of the conventional acoustic model in the speech recognition system of FIG. 30, and the data partitioning step of block 3310 is then repeated (block 3340 of FIG. 33). This may result in a change in the way in which Some of the acoustic observation vectors are bucketed, resulting in a more correct Set of training data. Then, using this improved training data set, the array of MBNs 3010 is re-trained using the training procedures described in this specification. This latter step is a repetition of the step of block 3330 of FIG.33. Thereafter, as more new acoustic observations are received, the bucketing Step of block 3310 is carried out for the incoming acoustic observation vectors, and the entire process of FIG.33 recycles in this manner." Thiesson at 38:53-39:16; *see also* Fig 33.  FIG. 33

Claim 14		
14.pre	A machine-readable storage, having stored thereon a computer program having a plurality of code sections executable by a machine for causing the machine to automatically generate from a first speech recognizer a second speech recognizer, said first speech recognizer comprising a first acoustic model with a first decision network and corresponding first phonetic contexts, and said second speech recognizer being adapted to a specific domain, said machine-readable storage causing the machine to perform the steps of:	*See* claim 1, including 1.pre.a-c. *See also*, "The advent of artificial intelligence within computer science has brought an abundance of decision-support systems. Decision-support systems are computer systems in which decisions, typically rendered by humans, are recommended and sometimes made. In creating decision-support systems, computer scientists seek to provide decisions with the greatest possible accuracy. Thus, computer scientists strive to create decision-support systems that are equivalent to or more accurate than a human expert. Applications of decision-support systems include medical diagnosis, troubleshooting computer networks, or other systems wherein a decision is based upon identifiable criteria." Thiesson at 1:18-29; *see* Thiesson claim 54. "With reference to FIG. 4, an exemplary system for implementing the invention includes a general purpose computing device in the form of a conventional personal computer 420, including a processing unit 421, a System memory 422, and a System buS 423 that couples various System components including the System memory to the processing unit 421. The system bus 423 may be any of several types of bus Structures including a memory bus or memory controller, a peripheral bus, and a local bus using any of a variety of bus architectures. The System memory includes read only memory (ROM) 424 and random access memory (RAM) 425. A basic input/output system 426 (BIOS), containing the basic process that helps to transfer information between elements within the personal computer 420, Such as during start-up, is stored in ROM 424. The personal computer 420 further includes a hard disk drive 427 for reading from and Writing to a hard disk, not shown, a magnetic disk drive 428 for reading from or writing to a removable magnetic disk 429, and an optical disk drive 430 for reading from or writing to a removable optical disk 431 Such as a CD ROM or other optical media. The hard disk drive 427, magnetic disk drive 428, and optical disk drive 430 are connected to the system bus 423 by a hard disk drive interface 432, a magnetic disk drive interface 433, and an optical drive interface 434, respectively. The drives and their associated computer-readable media provide nonvolatile Storage of computer readable instructions, data Structures, program modules and other data for the

		personal computer 420. Although the exemplary environment described herein employs a hard disk, a removable magnetic disk 429 and a removable optical disk 431, it should be appreciated by those skilled in the art that other types of computer readable media which can Store data that is accessible by a computer, Such as magnetic cassettes, flash memory cards, digital Video disks, Bernoulli cartridges, random access memories (RAMs), read only memories (ROM), and the like, may also be used in the exemplary operating environment. A number of program modules may be Stored on the hard disk, magnetic disk 429, optical disk 431, ROM 424 or RAM 425, including an operating system 435, one or more application programs 436, other program modules 437, and program data 438. A user may enter commands and information into the personal computer 420 through input devices such as a keyboard 440 and pointing device 442. Other input devices (not shown) may include a microphone, joystick, game pad, Satellite dish, Scanner, or the like. These and other input devices are often connected to the processing unit 421 through a Serial port interface 446 that is coupled to the System bus, but may be connected by other interfaces, Such as a parallel port, game port or a universal Serial bus (USB). A monitor 447 or other type of display device is also connected to the System bus 423 via an interface, Such as a video adapter 448. In addition to the monitor, personal computers typically include other peripheral output devices (not shown), Such as speakers and printers." Thiesson at 10:50-11:36.
14.a	based on said first acoustic model, generating a second acoustic model with a second decision network and corresponding second phonetic contexts for said second speech recognizer by re-estimating said first decision network and said corresponding first phonetic contexts based on domain-specific training data,	*See* claim 1, including 1.a.1-2, and 14.pre.

	wherein said first decision network and said second decision network utilize a phonetic decision tree to perform speech recognition operations,	
14.b	wherein the number of nodes in the second decision network is not fixed by the number of nodes in the first decision network, and wherein said re-estimating comprises partitioning said training data using said first decision network of said first speech recognizer.	*See* claim 1, including 1.b.1-2, and 14.pre, 14.a.
Claim 27		
27.pre	"A computerized method of generating a second speech recognizer comprising the steps of:"	*See* claim 1, including 1.pre.a-c.
27.a	"identifying a first speech recognizer of a first domain comprising a first acoustic model with a first decision network and corresponding first phonetic contexts;"	*See* claim 1, including 1.pre.b and 27.pre.
27.b	"receiving domain-specific training data of a second domain; and"	Thiesson discloses receiving domain-specific training data of a second domain. Thiesson discloses a fast bootstrapping procedure that inputs waveforms from the training data to

		partition the data and train the acoustic model. *See* claim 1 including 1.pre.c and 1.a.1. *See also*, "FIG. 33 illustrates a bootstrap training procedure involving the entire speech recognition system of FIG. 30 including the language model 3040. In this case, the training data consists of an input utterance waveform representing a predetermined (a priori known) sequence of senones. The speech recognition system of FIG. 30 partitions the data (block 3310 of FIG. 33) by assigning each successive acoustic observation vector to a particular senone "bucket' (block 3320 of FIG. 33). Initially, when this step is performed the speech recognition system of FIG. 30 can employ a conventional Structure as an acoustic model in place of the acoustic model 3010 consisting of the array of MBNs of the present invention. Once the 33-dimensional acoustic observation vectors have been assigned to various senone buckets by the step of block 3320, the resulting data is used to train the array of MBNs 3010 of FIG. 30 using the training procedures described in this specification (block 3330 of FIG. 33). The array of MBNs 3010 is then used as the acoustic model in place of the conventional acoustic model in the speech recognition system of FIG. 30, and the data partitioning step of block 3310 is then repeated (block 3340 of FIG. 33). This may result in a change in the way in which Some of the acoustic observation vectors are bucketed, resulting in a more correct Set of training data. Then, using this improved training data set, the array of MBNs 3010 is re-trained using the training procedures described in this specification. This latter step is a repetition of the step of block 3330 of FIG.33. Thereafter, as more new acoustic observations are received, the bucketing Step of block 3310 is carried out for the incoming acoustic observation vectors, and the entire process of FIG.33 recycles in this manner." Thiesson at 38:53-39:16; *see also* Thiesson at Fig. 33.
27.c	"based on the first speech recognizer and the domain-specific training data, generating a second acoustic model of said first domain and said second domain comprising a second	Thiesson in combination with the references mentioned in claim 1.a.1 disclose based on the first speech recognizer and the domain-specific training data, generating a second acoustic model of said first domain and said second domain comprising a second acoustic model with a second decision network and corresponding second phonetic contexts. This includes re-estimating the first decision network and first phonetic contexts. *See claim 1 including 1.pre.c and 1.a.1.*

	acoustic model with a second decision network and corresponding second phonetic contexts, wherein the first domain comprises at least a first language, wherein the second domain comprises at least a second language, and wherein the second speech recognizer is a multi-lingual speech recognizer.”	Moreover, the creation of a multilingual second recognizer was obvious in light of Thiesson in combination with any of the following references. For example, multilingual speech recognizers were well known at the time of the ’925 patent. *See, e.g.*, Schultz et al., “Polyphone Decision Tree Specialization for Language Adaptation”, ICASSP-2000, Istanbul, Turkey, Jun. 2000 (Section 2.2. describing generating and use of multilingual speech recognizers). The motivation to combine these references would at least include: • Combining prior art elements according to known methods to yield predictable results (multilingual recognizers were known using similar techniques for predictable results); • Simple substitution of one known element for another to obtain predictable results (substituting the domain from just one language to two); • Use of known technique to improve similar devices (methods, or products) in the same way (same technique of expanding domain of the recognizer to include multiple languages was known)); • Applying a known technique to a known device (method, or product) ready for improvement to yield predictable results (multilingual recognizers were known using similar known techniques for predictable results); • “Obvious to try”—choosing from a finite number of identified, predictable solutions, with a reasonable expectation of success (obvious to try to recognize more than one language); • Market forces and benefits associated with the known benefits of multilingual recognizers were predictable to POSA • Teaching of prior art would have lead a POSA to combine the references to arrive at a multilingual recognizer. For example, Sabourin discloses a multilingual system. Sabourin discloses a multilingual system where there are two different language domains as recited. It would be obvious to a person having ordinary skill in the art to combine Thiesson

		with Sabourin. Both Thiesson and Sabourin are in the same field of art. A person having ordinary skill in the art would be motivated to combine Thiesson with Sabourin. A person having ordinary skill in the art considering Thiesson's broad incorporation of unspecified training data would be motivated to seek out a system that builds a multilingual speech set, as disclosed by Sabourin. A person having ordinary skill in the art would have a reasonable expectation of success in implementing the teaching of Thiesson to generate a multilingual speech recognizer as provided by Sabourin because the combination involves the predictable use of prior art elements according to their established functions. "The invention relates to a method and apparatus for training a multilingual speech model Set. The multilingual Speech model Set generated is Suitable for use by a speech recognition System for recognizing spoken utterances for at least two different languages. The invention allows using a single Speech recognition unit with a single Speech model Set to perform Speech recognition on utterances from two or more languages. The method and apparatus make use of a group of a group of acoustic Sub-word units comprised of a first Subgroup of acoustic Sub-word units associated to a first language and a Second Subgroup of acoustic Sub-word units associated to a Second language where the first Subgroup and the Second Subgroup share at least one common acoustic Sub-word unit. The method and apparatus also make use of a plurality of letter to acoustic Sub-word unit rules Sets, each letter to acoustic Sub-word unit rules Set being associated to a different language. A Set of untrained speech models is trained on the basis of a training Set comprising speech tokens and their associated labels in combination with the group of acoustic Sub-word units and the plurality of letter to acoustic Sub-word unit rules Sets. The invention also provides a computer readable Storage medium comprising a program element for implementing the method for training a multilingual Speech model Set." Sabourin at Abstract. "A deficiency of the above-described method is that the Speech recognition System requires as an input the language associated to the input utterance, which may not be readily available to the Speech recognition System. Usually, obtaining the language requires prompting the user for the language of use thereby requiring an additionally Step in the Service being provided by the Speech recognition enabled system which may lower the level of satisfaction of the user with the system as a whole. Another deficiency of the

		above noted method is the costs associated to developing and maintaining a Speech model Set for each language the Speech recognition System is adapted to recognize. More Specifically, each speech model Set must be trained individually, a task requiring manpower for each individual language thereby increasing significantly the cost of Speech recognition Systems operating in multilingual environments with respect to Systems operating in unilingual environments. In addition, the above-described method requires the Storage of a speech model Set for each language in memory thereby increasing the cost of the Speech recognition System in terms of memory requirements. Finally, the above described method requires testing a speech model Set for each language thereby increasing the testing cost of the Speech recognition System for each language the Speech recognition System is adapted to recognize. Thus, there exists a need in the industry to refine the process of training Speech models So as to obtain an improved multilingual Speech model Set capable of being used by a speech recognition System for recognizing spoken utterances for at least two different languages." Sabourin at 1:55-2:17. "In accordance with another broad aspect, the invention provides a method for generating a multilingual speech model Set Suitable for use in a multilingual speech recognition System. The method comprises providing a group of acoustic Sub-word units having a first Subgroup of acoustic Sub-word units associated to a first language and a Second Subgroup of acoustic Sub-word units associated to a Second language. The first Subgroup and the Second Subgroup share at least one common acoustic Sub-word unit. The method further comprises providing a training Set comprising a plurality of entries, each entry having a speech token representative of a word and a label being an orthographic representation of the word. The method further comprises providing a Set of untrained speech models and training the Set of untrained speech models by utilizing the training Set, the plurality of letter to acoustic Sub-word unit rules Sets and the group of acoustic Sub-word units to derive the multilingual Speech model Set." Sabourin at 2:57-3:7.

		"The invention provides a method for generating the multilingual speech model set shown in FIG.1. As shown in FIG. 2 of the drawings, the method comprises providing 200 a group of acoustic Sub-word units comprised of a first Subgroup of acoustic Sub-word units associated to a first language and a Second Subgroup of acoustic Sub-word units associated to a Second language. The first Subgroup and the Second Subgroup share at least one common acoustic Sub word unit. In a preferred embodiment, the group of acoustic Sub word unit is obtained by combining sub-word units from individual languages. For each language for which the multilingual Speech model Set is to be representative of, an inventory of sub-word units can be obtained. The acoustic Sub-word units are basic units of Sound in a language. These inventories may be produced by phoneticians or may be provided as off-the-shelf components. In a specific example, the Sub-word units are phonemes. The Skilled person in the art will readily see that other sub-word units may be used here without detracting from the spirit of the invention." Sabourin at 4:25-45. "The training of the set of untrained speech models further comprises processing 304 the group of transcriptions generated at step 300 on the basis of a speech token of the corresponding entry in the training Set whereby training the Set of untrained speech models to derive the multilingual Speech model Set." Sabourin at 9:34-39. *See* Figs. 1-7.

APPENDIX A-2
Invalidity Claim Chart for U.S. Pat. No. 6,999,925 ('925 patent)
U.S. Pat. No. 6,912,499, Sabourin et al. ("Sabourin")[1]

On October 16, 2020, Nuance narrowed the asserted '925 patent claims to 1, 14 and 27. Nuance's Disclosure Limiting Asserted Claims of Phase 1 Patents. Claims 1, 14 and 27 of the '925 patent are anticipated and/or rendered obvious by Sabourin alone or in combination with at least the following references. The chart below discloses how prior art references identified by Omilia disclose, either expressly or inherently, and/or render obvious each of the asserted claims:

1. R. Singh, et al., *Domain Adduced State Tying For Cross-Domain Acoustic Modeling*, Proc. Of the 6th Europ. Conf. on Speech Communication and Technology, Budapest (1999).[2]

Sabourin incorporates by reference the following references:

1. U.S. Pat. No. 5,195,167 by Bahl et al., "Apparatus and Method of Grouping Utterance of a Phoneme into Context-Dependent Categories based on Sound-Similarity for Automatic Speech Recognition", Mar. 16, 1993.

This invalidity claim chart is based on Omilia's present understanding of the asserted claims, the Court's August 6, 2020 Claim Construction Order (ECF. No. 157), Nuance's initial and all subsequent supplemental Infringement Contentions, its July 7, 2020 Response to Omilia's Supplemental Non-Infringement and Invalidity Responses, Nuance's Response to Omilia's Interrogatory No. 9, and Omilia's investigation to date. Omilia is not adopting Nuance's apparent constructions, nor is Omilia admitting to the accuracy of any particular construction. Omilia reserves all rights to amend this invalidity claim chart if Nuance further amends its Infringement Contentions, or in response to any statements made by Nuance at any phase of this litigation. In addition, Discovery is still ongoing, and Omilia reserves its right to supplement its contentions with additional evidence as discovery continues.

The citations provided are exemplary and do not necessarily include each and every disclosure of the limitation in the references. Omilia has cited the most relevant portions of the identified prior art. Other portions of the identified prior art may additionally disclose, either expressly or inherently, and/or render obvious one or more limitations of the asserted claims. To provide context or aid in understanding the prior art and the state of the art at the time of the alleged invention, Omilia reserves the right to rely on: (1) uncited portions of the identified prior art; (2) other prior art not identified herein; (3) references that show the state of the art (irrespective of whether such

---

[1] Sabourin was filed on August 31, 1999 and is prior art at least under 35 §102(a) & 102(e).

[2] Singh was published at the conference in September 5-9, 1999 and is prior art at least under 35 §102(a) & 102(b).

references themselves qualify as prior art to the asserted patents); (4) factual testimony from the inventors or authors of the prior art references, or purveyors of prior art devices; and/or (5) expert testimony.

Citations to a particular drawing or figure in the accompanying charts should be understood to include the description of the drawing or figure, as well as any text associated with the drawing or figure. Citations to particular text concerning a drawing or figure, should also be understood to encompasses that drawing or figure as well. The identified prior art expressly and/or inherently discloses the features of the asserted claims under all proposed constructions of the asserted claims. To establish the inherency of certain features of the prior art to invalidate the asserted claims, Omilia may rely on cited or uncited portions of the prior art, other documents, factual testimony, and expert testimony. To the extent that the Court finds that this reference does not explicitly teach certain limitations in the asserted claims, such limitations would have been inherent and/or obvious and Omilia reserves the right to supplement or amend this claim chart to address such a finding.

’925 Patent		
Claim 1		
1.pre.a	“A computerized method of automatically generating from a first speech recognizer a second speech recognizer”	Sabourin discloses automatically generating a second speech recognizer from a first speech recognizer. Sabourin describes using a computer method to create a multilingual speech recognizer based on other speech recognizers. Sabourin discloses at least a set of speech models of a first language to generate speech models of a second language. *See, e.g.*: “In another embodiment, the labels can be assigned using a speech recognizer unit by Selecting the top scoring recognition candidate as the label for the utterance processed by the recognizer.” Sabourin at 6:12-15. “This invention relates to speech model sets and to a method and apparatus for training speech model sets for use in speech recognition systems operating in multilingual environments as may be used in a telephone directory assistance system, voice activated dialing (VAD) system, personal voice dialing systems and other speech recognition enabled services.”

’925 Patent		
		Sabourin at 1:6-13; *see also* Title (54) (“Method and Apparatus for Training A Multilingual Speech Model Set”). “The invention relates to a method and apparatus for training a multilingual speech model Set. The multilingual Speech model Set generated is Suitable for use by a speech recognition System for recognizing spoken utterances for at least two different languages. The invention allows using a single Speech recognition unit with a single Speech model Set to perform Speech recognition on utterances from two or more languages. The method and apparatus make use of a group of a group of acoustic Sub-word units comprised of a first Subgroup of acoustic Sub-word units associated to a first language and a Second Subgroup of acoustic Sub-word units associated to a Second language where the first Subgroup and the Second Subgroup share at least one common acoustic Sub-word unit. The method and apparatus also make use of a plurality of letter to acoustic Sub-word unit rules Sets, each letter to acoustic Sub-word unit rules Set being associated to a different language. A Set of untrained speech models is trained on the basis of a training Set comprising speech tokens and their associated labels in combination with the group of acoustic Sub-word units and the plurality of letter to acoustic Sub-word unit rules Sets. The invention also provides a computer readable Storage medium comprising a program element for implementing the method for training a multilingual Speech model Set.” Sabourin at Abstract (57). “The invention provides a method for generating the multilingual speech model set shown in FIG.1. As shown in FIG. 2 of the drawings, the method comprises providing 200 a group of acoustic Sub-word units comprised of a first Subgroup of acoustic Sub-word units associated to a first language and a Second Subgroup of acoustic Sub-word units associated to a Second language. The first Subgroup and the Second Subgroup share at least one common acoustic Sub word unit. In a preferred embodiment, the group of acoustic Sub word unit is obtained by combining sub-word units from individual languages. For each language for which the multilingual Speech model Set is to be representative of, an inventory of sub-word units can be obtained. The acoustic Sub-

’925 Patent		
		word units are basic units of Sound in a language. These inventories may be produced by phoneticians or may be provided as off-the-shelf components. In a specific example, the Sub-word units are phonemes. The Skilled person in the art will readily see that other sub-word units may be used here without detracting from the spirit of the invention.” Sabourin at 4:24-45. “The invention provides a method for initializing a speech model set for a first language on the basis of a speech model set from a second language different from the first language. In the preferred embodiment, the invention makes use of the feature descriptions of the sub-word unit to generate initialization data for the speech models. In specific example, suppose we have a first language for which speech models are available and a second language for which speech models are not available. In addition, suppose that in the second language, there is an acoustic sub-word unit, herein designated as the new acoustic sub-word unit, that is not comprised in the acoustic sub-word inventory of the first language. The acoustic sub-word units common to the first language and the second language are initialized with the speech models associated to the first language. This invention provides a method for initializing the speech model of a new acoustic sub-word unit on the basis of the known speech models associated to the first language more specifically by using the nearest phoneme as a basis to initialization. For example, say the nearest phoneme to /X/ according to a certain criteria is phoneme /Y/ and that the speech model for /Y/ is known. Initialization involves copying all the model parameters (eg. State transition weights) for model /Y/ into a model for /X/. This is particularly advantageous for initializing the speech model for sub-word units in a language for speech models are not available. The method will be described below for acoustic sub-word units being phonemes. The skilled person in the art will readily observe that this method may be applied to other types or acoustic sub-word units without detracting from the spirit of the invention.” Sabourin at 6:45-7:8. Claim 1 (“a method for generating a multilingual speech model set, . . . providing a set of untrained speech model, said set of untrained speech models have at least a first

’925 Patent		
		untrained speech model . . . training a set of untrained speech models by utilizing said training set, a plurality of letter to acoustic sub-word unit rules sets and said group of acoustic sub-word units to derive the multilingual speech model set, …”). *See also* other claims (e.g. claim 11-24). Sabourin Claim 1 “The above-described method for generating a multilingual speech model Set can also be implemented on any suitable computing platform as shown in FIG. 6. Such computing platform typically includes a CPU 602 and a memory or computer readable medium 600 connected to the CPU 602 by a data communication bus. The memory stores the data 606 and the instructions of the program element 604 implementing the functional blocks depicted in the drawings and described in the Specification. In a specific example, the program element 604 implements the processing unit 408 and the data 606 comprises the group of acoustic Sub-word units, the plurality of letter to acoustic Sub-word units rules Sets, the training Sets and the untrained Speech models. The program element 604 operates on the data 606 in accordance with the algorithms described above to generate a multilingual Speech model Set using the techniques described in this specification.” Sabourin at 13:48-64 Sabourin at Figs. 1-5, 7.
1.pre.b	“said first speech recognizer comprising a first acoustic model with a first decisions network and corresponding first phonetic contexts”	Sabourin discloses a first speech recognizer comprising a first acoustic model with a first decision network and corresponding first phonetic contexts. Sabourin discloses the speech models that have corresponding phonetic contexts of phonemes. In addition, Sabourin describes the use of HMMs, which make use of decision networks with corresponding phonetic contexts. *See, e.g.,* “In a preferred embodiment, as shown in FIG. 1, the invention provides a computer readable Storage medium comprising a data Structure containing a multilingual speech model set 100. The multilingual speech model set 100 is Suitable for use in a speech

’925 Patent		
		recognition System for recognizing spoken utterances for at least two different languages. The multilingual speech model Set comprises a first Subset of Speech models associated to a first language and a Second Subset of Speech models associated to a Second language. The first Subset and the Second Subset share at least one common Speech model. Preferably, a single copy of the shared common Speech model is Stored on the computer readable medium. The data Structure containing a multilingual Speech model Set 100 provides an association between the speech models in the multilingual speech model set 100 and their respective acoustic Sub-word unit. In a specific example, the acoustic Sub-word units are phonemes. Optionally, the Speech models in the Speech model Set maybe representative of the allophonic context of the phonemes. In these cases, the data Structure containing a multilingual speech model set 100 provides an association between the speech models in the multilingual speech model set 100 and their respective allophones.” Sabourin at 3:52-4:7. “In a preferred embodiment, the first subset and the second subset share a plurality of speech models. In a specific example the speech models in the multilingual speech models set 100 are represented by Hidden Markov Models. Hidden Markov Models are well known in the art to which this invention pertains and will not be described in further detail here.” Sabourin at 4:16-23. “In a preferred embodiment, the group of acoustic Sub word unit is obtained by combining sub-word units from individual languages. For each language for which the multilingual Speech model Set is to be representative of, an inventory of sub-word units can be obtained. The acoustic Sub-word units are basic units of Sound in a language. These inventories may be produced by phoneticians or may be provided as off-the-shelf components. In a specific example, the Sub-word units are phonemes. The Skilled person in the art will readily see that other sub-word units may be used here without

’925 Patent		
		detracting from the spirit of the invention. The table below shows an example of a phoneme inventory for the Spanish language. … Typically each language possesses a unique inventory of acoustic Sub-word units. Preferably, each acoustic Sub-word unit in the acoustic Sub-word unit inventory for each language is associated to a feature description. The feature description provides a convenient means for establishing confusability rules for a given language and provide initialization information for the recognizer. The table below shows an example of feature descriptions of a Subset of the phonemes for the Spanish language.” Sabourin at 4:36-67. “The method also comprises providing 206 a set of untrained speech models. Each speech model is associated to an acoustic sub-word unit in the group of acoustic sub-word units. In a specific example, the untrained speech models are HMMs and are assigned a complex topology of about 80 means per phonemes and three HMMs states. The untrained speech models may be obtained as off-the-shelf components or may be generated using a nearest sub-word unit method.” Sabourin at 6:16-24. “In a preferred embodiment, the Speech models are trained using a maximum likelihood method. The speech models are HMMs having a set of States and transitions between these States, each State represented by a State probability and a set of transition probabilities. The state probability is modeled as a mixture of Gaussian distributions where each Gaussian distribution has a means and a covariance matrix. The means and covariance matrices may be shared be other models without detracting from the spirit of the invention. The transition probabilities are exponential distributions with mean u. In a Specific example, the untrained speech models are trained using maximum a posteriori adaptation (MAP) as described in Gauvain J.-L. and Lee C. -H. (1994) “maxi mum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains'

’925 Patent		
		IEEE Trans. Speech Audio Process. 2, pp. 291-298. The content of this document is hereby incorporated by reference. Other methods of training Speech models on the basis of Speech tokens may be used here without detracting from the Spirit of the invention. In a preferred embodiment Several training iterations are performed to condition the Speech models.” Sabourin at 9:55-10:20. “In another specific example, the acoustic Sub-word units are Selected from the Set consisting of allophones, triphones biphones or any other representation of a phoneme on the basis of context.” Sabourin at 10:24-27 “Having generated multilingual Speech models associated to phonemes, the Speech labels in the training Set are Segmented on the basis of the phonemes and more precise Speech models are build. The allophonic context is defined on the basis of adjoining phonemes and an allophonic decision tree.” Sabourin at 10:37-41.
1.pre.c	“and said second speech recognizer being adapted to a specific domain, said method comprising:”	Sabourin discloses that the second speech recognizer (which is multilingual) is adapted for a specific domain. For example, Sabourin discloses initializing a speech model set for sub-word units in the second language that are not present in the first language. *See, e.g.*, “The invention relates to a method and apparatus for training a multilingual speech model Set. The multilingual Speech model Set generated is Suitable for use by a speech recognition System for recognizing spoken utterances for at least two different languages. The invention allows using a single Speech recognition unit with a single Speech model Set to perform Speech recognition on utterances from two or more languages. The method and apparatus make use of a group of a group of acoustic Sub-

’925 Patent		
		word units comprised of a first Subgroup of acoustic Sub-word units associated to a first language and a Second Subgroup of acoustic Sub-word units associated to a Second language where the first Subgroup and the Second Subgroup share at least one common acoustic Sub-word unit. The method and apparatus also make use of a plurality of letter to acoustic Sub-word unit rules Sets, each letter to acoustic Sub-word unit rules Set being associated to a different language. A Set of untrained speech models is trained on the basis of a training Set comprising speech tokens and their associated labels in combination with the group of acoustic Sub-word units and the plurality of letter to acoustic Sub-word unit rules Sets. The invention also provides a computer readable Storage medium comprising a program element for implementing the method for training a multilingual Speech model Set.” Sabourin at Abstract. “In accordance with a broad aspect, the invention provides a computer readable storage medium having a data structure containing a multilingual speech model set. The multilingual speech model set is suitable for use in a speech recognition system for recognizing spoken utterances for at least two different languages.” Sabourin at 2:23-28. “The invention provides a method for initializing a speech model set for a first language on the basis of a speech model set form a second language different from the first language. In the preferred embodiment, the invention makes use of the feature descriptions of the sub-word unit to generate initialization data for the speech models. In specific example, suppose we have a first language for which speech models are available and a second language for which speech models are not available. In addition, suppose that in the second language, there is an acoustic sub-word unit, herein designated as the new acoustic sub-word unit, that is not comprised in the acoustic sub-word inventory of the first language. The acoustic sub-word units common to the first language and the second language are initialized with the speech models associated to the first language. This invention provides a method for initializing the speech model of new acoustic sub-word unit on the basis of the known

’925 Patent		
		speech models associated to the first language more specifically by using the nearest phoneme as a basis to initialization. For example, say the nearest phoneme to /X/ according to a certain criteria is phoneme /Y/ and that the speech model for /Y/ is known. Initialization involves copying all the model parameters (eg. State transition weights) for model /Y/ into a model for /X/. This is particularly advantageous for initializing the speech model for sub-word units in a language for speech models are not available. The method will be described below for acoustic sub-word units being phonemes. The skilled person in the art will readily observe that this method may be applied to other types or acoustic sub-word units without detracting from the spirit of the invention.” Sabourin at 6:45-7:8 Sabourin at Figs. 1-5, 7. *See, e.g.*, Sabourin at claim 1 (other claims as well)
1.a.1	“based on said first acoustic model, generating a second acoustic model with a second decision network and corresponding second phonetic contexts for said second speech recognizer by re-estimating said first decision network and said corresponding first phonetic contexts based on domain-specific training data,”	Sabourin discloses based on said first acoustic model, generating a second acoustic model with a second decision network and corresponding second phonetic contexts for said second speech recognizer by re-estimating said first decision network and said corresponding first phonetic contexts based on domain-specific training data. In Sabourin, a first speech model is adapted based on acoustic information to create a second multilingual speech model with different phonetic contexts than the first. For example, Sabourin adapts speech models of a first language to new sub-word units in the second language that are not present in the first language. *See, e.g.,:* “The method also comprises providing 204 a training Set comprising a plurality of entries, each entry having a speech token representative of a Vocabulary item and a label being an orthographic representation of the Vocabulary item. The training Set may be obtained by Storing speech tokens and manually writing out each Vocabulary item associated to the Speech token or by having individuals read a predetermined Sequence of Vocabulary items. Such training Sets are also available as off-the-shelf components. In a specific example of implementation, the training Sets are reasonably accurate in that

’925 Patent		
		the Speech tokens have a high likelihood of corresponding to the written vocabulary items. A training Set is provided for each language that the multilingual speech model Set is to be representative of. In another embodiment, the labels can be assigned using a speech recognizer unit by Selecting the top scoring recognition candidate as the label for the utterance processed by the recognizer. The method also comprises providing 206 a set of untrained speech models. Each speech model is associated to an acoustic Sub-word unit in the group of acoustic Sub-word units. In a specific example, the untrained speech models are HMMS and are assigned a complex topology of about 80 means per phonemes and three HMMs states. The untrained Speech models may be obtained as off-the-shelf components or may be generated by using a nearest Sub word unit method. In the Specific example where the Sub word units are phonemes, a nearest phoneme method is used. The nearest phoneme method is described in detail below. The skilled person in the art will readily observe that it may be applied to acoustic Sub-word units other than phonemes without detracting from the Spirit of the invention. In its broad aspect, a nearest phoneme method includes initializing a speech model of a new phoneme on the basis of a speech model of a phoneme that is acoustically similar to the new phoneme. The phoneme acoustically Similar to the new phoneme may be determined by a human which determines the closest Sounding phoneme or by a heuristic algorithm implemented on a computer readable medium. In the preferred embodiment, the nearest phoneme is derived on the basis of a phonological Similarity method. Alternatively, the Speech models may be initialized by assigning random values to the initial models and by hand aligning the Speech tokens with their corresponding transcriptions for a large number of Speech tokens and training the initial model values. This alternative method is particularly useful when there are no speech models available.” Sabourin at 5:66-6:44 “The invention provides a method for initializing a speech model set for a first language on the basis of a speech model set form a second language different from the first language. In the preferred embodiment, the invention makes use of the feature

’925 Patent		
		descriptions of the sub-word unit to generate initialization data for the speech models. In specific example, suppose we have a first language for which speech models are available and a second language for which speech models are not available. In addition, suppose that in the second language, there is an acoustic sub-word unit, herein designated as the new acoustic sub-word unit, that is not comprised in the acoustic sub-word inventory of the first language. The acoustic sub-word units common to the first language and the second language are initialized with the speech models associated to the first language. This invention provides a method for initializing the speech model of new acoustic sub-word unit on the basis of the known speech models associated to the first language more specifically by using the nearest phoneme as a basis to initialization. For example, say the nearest phoneme to /X/ according to a certain criteria is phoneme /Y/ and that the speech model for /Y/ is known. Initialization involves copying all the model parameters (e.g. State transition weights) for model /Y/ into a model for /X/. This is particularly advantageous for initializing the speech model for sub-word units in a language for speech models are not available. The method will be described below for acoustic sub-word units being phonemes. The skilled person in the art will readily observe that this method may be applied to other types or acoustic sub-word units without detracting from the spirit of the invention.” Sabourin at 6:45-7:8. “The invention relates to a method and apparatus for training a multilingual speech model Set. The multilingual Speech model Set generated is Suitable for use by a speech recognition System for recognizing spoken utterances for at least two different languages. The invention allows using a single Speech recognition unit with a single Speech model Set to perform Speech recognition on utterances from two or more languages. The method and apparatus make use of a group of a group of acoustic Sub-word units comprised of a first Subgroup of acoustic Sub-word units associated to a first language and a Second Subgroup of acoustic Sub-word units associated to a Second language where the first Subgroup and the Second Subgroup share at least one common acoustic Sub-word unit. The method and apparatus also make use of a plurality of letter

’925 Patent		
		to acoustic Sub-word unit rules Sets, each letter to acoustic Sub-word unit rules Set being associated to a different language. A Set of untrained speech models is trained on the basis of a training Set comprising speech tokens and their associated labels in combination with the group of acoustic Sub-word units and the plurality of letter to acoustic Sub-word unit rules Sets. The invention also provides a computer readable Storage medium comprising a program element for implementing the method for training a multilingual Speech model Set.” Sabourin at Abstract. “In accordance with another broad aspect, the invention provides a method for generating a multilingual speech model Set Suitable for use in a multilingual speech recognition System. The method comprises providing a group of acoustic Sub-word units having a first Subgroup of acoustic Sub-word units associated to a first language and a Second Subgroup of acoustic Sub-word units associated to a Second language. The first Subgroup and the Second Subgroup share at least one common acoustic Sub-word unit. The method further comprises providing a training Set comprising a plurality of entries, each entry having a speech token representative of a word and a label being an orthographic representation of the word. The method further comprises providing a Set of untrained speech models and training the Set of untrained speech models by utilizing the training Set, the plurality of letter to acoustic Sub-word unit rules Sets and the group of acoustic Sub-word units to derive the multilingual Speech model Set.” Sabourin at 2:57-3:6. “In a preferred embodiment, as shown in FIG. 1, the invention provides a computer readable Storage medium comprising a data Structure containing a multilingual speech model set 100. The multilingual speech model set 100 is Suitable for use in a speech recognition System for recognizing spoken utterances for at least two different languages. The multilingual speech model Set comprises a first Subset of Speech models associated to a first language and a Second Subset of Speech models associated to a Second language. The first Subset and the Second Subset share at least one common Speech model. Preferably, a single copy of the shared common Speech model

’925 Patent		
		is Stored on the computer readable medium. The data Structure containing a multilingual Speech model Set 100 provides an association between the speech models in the multilingual speech model set 100 and their respective acoustic Sub-word unit. In a specific example, the acoustic Sub-word units are phonemes. Optionally, the Speech models in the Speech model Set maybe representative of the allophonic context of the phonemes. In these cases, the data Structure containing a multilingual speech model set 100 provides an association between the speech models in the multilingual speech model set 100 and their respective allophones.” Sabourin at 3:52-4:7. “In a preferred embodiment, the first subset and the second subset share a plurality of speech models. In a specific example the speech models in the multilingual speech models set 100 are represented by Hidden Markov Models. Hidden Markov Models are well known in the art to which this invention pertains and will not be described in further detail here. The invention provides a method for generating the multilingual speech model set shown in FIG.1. As shown in FIG. 2 of the drawings, the method comprises providing 200 a group of acoustic Sub-word units comprised of a first Subgroup of acoustic Sub-word units associated to a first language and a Second Subgroup of acoustic Sub-word units associated to a Second language. The first Subgroup and the Second Subgroup share at least one common acoustic Sub word unit. In a preferred embodiment, the group of acoustic Sub word unit is obtained by combining sub-word units from individual languages. For each language for which the multilingual Speech model Set is to be representative of, an inventory of sub-word units can be obtained. The acoustic Sub-word units are basic units of Sound in a language. These inventories may be produced by phoneticians or may be provided as off-the-shelf components. In a specific example, the Sub-word units are phonemes. The Skilled person in the art will readily see that other sub-word units may be used here without detracting from the spirit of the invention.”

’925 Patent		
		Sabourin at 4:16-45. “The method also comprises training 208 the set of untrained Speech models by utilizing the training Set, the plurality of letter to acoustic Sub-word unit rules Sets and the group of acoustic Sub-word units to derive the multilingual speech model set 100. In a preferred embodiment, as shown in FIG. 3 of the drawings, training 208 the Set of untrained speech models comprises processing 300 at least Some of the entries in the training Set on the basis of a certain letter to acoustic Sub-word unit rules Set in the plurality of letter to acoustic Sub-word unit rules Sets to derive a group of transcriptions. In a Specific example, there is a plurality of training Sets, each Set being associated to a language. A given training Set associated to a given language is first processed to extract the labels from each entry. This operation will be readily available to those skilled in the art. The labels extracted from the given training Set are then processed by a given letter to acoustic Sub-word unit rules Set, the given letter to acoustic Sub-word unit rules Set being associated to the given language. The result of this processing is a group of transcriptions associated to a given language, the transcriptions in the group of transcriptions comprising a Sequence of acoustic Sub-word units of the group of acoustic Sub-word units. This processing is performed on all the training Sets of the plurality of training Sets. Optionally, the groups of transcriptions from each respective training Set are combined to form a compound training Set. As a variant, the training Set comprises a plurality of entries, the entries being representative of Vocabulary items in one or more languages. The training Set is first processed to extract the labels from each entry. The labels extracted from the training Set are then processed by a Subset of the plurality of letter to acoustic Sub-word unit rules Sets to derive a plurality of transcriptions. A first transcription in the plurality of transcriptions corresponding to a certain vocabulary item is associated to a first language and a Second transcription in the plurality of transcriptions corresponding to the same certain vocabulary item is associated to a Second language different from the first language. As yet another variant, prosodic rules may be used to process the transcriptions generated by with the letter to acoustic unit rules Sets to derive Surface form transcriptions. The prosodic rules are associated to respective languages may be applied

’925 Patent		
		to the groups of transcriptions associated to the same languages. Alternatively, a Subset of the prosodic rules may be applies to the transcriptions associated to languages different from the languages associated to the prosodic rules being considered. In accordance with a preferred embodiment, training the Set of untrained Speech models further comprises associating 302 the acoustic Sub-word units in the group of acoustic Sub-word units to respective speech models in the Set of untrained speech models. The training of the Set of untrained speech models further comprises processing 304 the group of transcriptions generated at step 300 on the basis of a speech token of the corresponding entry in the training Set whereby training the Set of untrained speech models to derive the multilingual Speech model Set. In a specific example of implementation, the Speech tokens in Said training Set are processed by a feature extraction unit to convert the Speech Signals into a Set of numeric values for Storage in a vector representative of acoustic information. A specific example of a vector is a cepstral vector. Each speech token is associated to a set of cepstral vectors, one vector being associated to a frame of the Speech token. The cepstral vectors are then aligned with the phonemes of the transcription corresponding to the Speech token. In a specific example, for the initial alignment, the cepstral vectors of the Speech token are randomly divided between the phonemes of the transcription. In another specific example, vector quantization or Kohonen associative mapping is used to provide the initial alignment of the cepstral vectors and the phonemes. These cepstral vectors are then used to condition the untrained speech models thereby training the Speech models. In a preferred embodiment, the Speech models are trained using a maximum likelihood method. The speech models are HMMs having a set of States and transitions between these States, each State represented by a State probability and a set of transition probabilities. The state probability is modeled as a mixture of Gaussian distributions where each Gaussian distribution has a means and a covariance matrix. The means and covariance matrices may be shared be other models without detracting from the spirit of the invention. The transition probabilities are exponential distributions with mean u. In a Specific example, the untrained speech models are trained using maximum a posteriori adaptation (MAP) as described in Gauvain J.-L. and Lee C. -H. (1994) “maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains' IEEE Trans. Speech

’925 Patent		
		Audio Process. 2, pp. 291-298. The content of this document is hereby incorporated by reference. Other methods of training Speech models on the basis of Speech tokens may be used here without detracting from the Spirit of the invention. In a preferred embodiment Several training iterations are performed to condition the Speech models. The iterations are performed by obtaining another assignation of the vectors to the phonemic transcription of the Speech label and restarting the training.” Sabourin at 8:49-10:23. Sabourin at Figs. 1-5, 7. *See, e.g.*, Sabourin, claim 1 (and other claims). *See also 1.pre.b* (discussing acoustic models, decision networks and phonetic context)
1.a.2	“wherein said first decision network and said second decision network utilize a phonetic decision free [sic] to perform speech recognition operations”	Sabourin discloses that the first decision network and second decision network utilize a phonetic decision tree in speech recognition operations. *See, e.g.*, “The allophonic context is defined on the basis of adjoining phonemes and an allophonic decision tree. A question set is provided to build the decision tree allophones. Decision trees have been described in U.S. Pat. No. 5,195,167 by Bahl et al. “Apparatus and Method of Grouping Utterance of a Phoneme into Context-Dependent Categories based on Sound-Similarity for Automatic Speech Recognition”, Mar. 16, 1993. The content of this document is hereby incorporated by reference. The invention provides a universal question set by grouping acoustic sub-word units into classes and develop a question set on the basis of the class and feature scores. Typically, each question involves a class of phonemes ranging from single phonemes to substantially large sets. Larger classes are built by concatenating smaller classes. The table below shows a specific example of entries in the universal question set. In a specific example of implementation, the universal question set is language independent.” Sabourin at 10:39-57.

’925 Patent		
		“The method also comprises providing 206 a set of untrained speech models. Each speech model is associated to an acoustic sub-word unit in the group of acoustic sub-word units. In a specific example, the untrained speech models are HMMs and are assigned a complex topology of about 80 means per phonemes and three HMMs states. The untrained speech models may be obtained as off-the-shelf components or may be generated using a nearest sub-word unit method.” Sabourin at 6:16-24. “In a preferred embodiment, the Speech models are trained using a maximum likelihood method. The speech models are HMMs having a set of States and transitions between these States, each State represented by a State probability and a set of transition probabilities. The state probability is modeled as a mixture of Gaussian distributions where each Gaussian distribution has a means and a covariance matrix. The means and covariance matrices may be shared be other models without detracting from the spirit of the invention. The transition probabilities are exponential distributions with mean u. In a Specific example, the untrained speech models are trained using maximum a posteriori adaptation (MAP) as described in Gauvain J.-L. and Lee C. -H. (1994) “maxi mum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains' IEEE Trans. Speech Audio Process. 2, pp. 291-298. The content of this document is hereby incorporated by reference. Other methods of training Speech models on the basis of Speech tokens may be used here without detracting from the Spirit of the invention. In a preferred embodiment Several training iterations are performed to condition the Speech models.” Sabourin at 9:55-10:20. “Once the allophonic context is defined, the allophones are extracted by pruning the allophonic decision tree where appropriate. For each Speech token, the corresponding phonemic transcription is examined. For each phoneme in the transcription, the adjacent “N” phoneme neighbors are examined for membership in a class of phonemes. The classes in which the neighboring phonemes belong are tabulated. The phonemic

’925 Patent		
		transcriptions associated to the Speech labels of the training Set are relabeled with the allophones on the basis of the defined allophonic context.” Sabourin at 11:23-32. Second, given this disclosure, use of phonetic decision trees to recognize speech was obvious in light of Sabourin in combination with Singh. The motivation to combine these references would at least include: • Combining prior art elements according to known methods to yield predictable results (use of phonetic decision trees were utilized in speech recognition using similar techniques for predictable results); • Simple substitution of one known element for another to obtain predictable results (substituting the type of decision network to a decision tree is simple and predicable); • Use of known technique to improve similar devices (methods, or products) in the same way (same technique of adapting the decision tree in a recognizer was known in speech recognizers)); • Applying a known technique to a known device (method, or product) ready for improvement to yield predictable results (use of decision trees in speech recognition was known using similar known techniques for predictable results); • “Obvious to try”—choosing from a finite number of identified, predictable solutions, with a reasonable expectation of success (obvious to try to use decision trees in speech recognition); • Market forces and benefits associated with the known benefits of using decision trees were predictable to POSA

’925 Patent		
		• Teaching of prior art would have lead a POSA to combine the references to arrive at a speech recognizer that uses decision trees. For example, Singh discloses that HMMs are organized as decision trees that are used to recognize speech. It would be obvious to a person having ordinary skill in the art to combine Sabourin with Singh to disclose adding nodes, pruning nodes and merging nodes in the decision network. Both Sabourin and Singh are in the same field of art. A person having ordinary skill in the art would be motivated to combine Sabourin with Singh. A person having ordinary skill in the art considering Sabourin’s disclosure that the HMMs “may be off-the-shelf components,” would be motivated to seek out a system or off-the-shelf component that adapts that explicitly organizes HMMs as decision trees, as disclosed by Singh, to further utilize the HMM tree structure. Further, Sabourin’s incorporation by reference of Bahl describing decision trees, suggests the potential use of trees with Sabourin. A person having ordinary skill in the art would have a reasonable expectation of success in implementing the teaching of Sabourin to organize the HMM in phonetic decision trees and use the trees to recognize speech as provided by Singh, because the combination involves the predictable use of prior art elements according to their established functions. Sabourin at 10:42-47: “A question set is provided to build the decision tree allophones. Decision tress have been described in U.S. Pat. No. 5,195,167 by Bahl et al. “Apparatus and Method of Grouping Utterance of a Phoneme into Context-Dependent Categories based on Sound-Similarity for Automatic Speech Recognition,” Mar. 16, 1993.” “This reduced set of parameters is obtained by grouping triphones into a statistically estimable number of clusters using decision trees [3]. For ASR systems based on Hidden Markov Models (HMMs), the decision trees result in sharing of output probability distribution functions across states, a procedure well known as state tying. Canonically, parameters are distributed (i.e. the states are tied) so as to best capture the acoustic-phonetic structure of the training corpus. In CDM, however, the acoustic-phonetic structure of the task domain may be significantly different from that of the training

’925 Patent		
		corpus. This disparity alone largely accounts for the degradation in performance when ASR systems are trained from out-of-domain corpora. In this paper we explore the possibility of compensating for some of this disparity by redistributing the states of HMM-based ASR systems according to the acoustic phonetic structure of the task domain data, rather than that of the training domain data.” Singh at Section 1.
1.b.1	“wherein the number of nodes in the second decision network is not fixed by the number of nodes in the first decision network,”	Sabourin discloses that “the number of nodes in the second decision network is not fixed by the number of nodes in the first decision network.” For example, Sabourin describes copying those subword units common between the two languages and adding subword units for the second language not present in the first language. *See. e.g.,:* “The invention relates to a method and apparatus for training a multilingual speech model Set. The multilingual Speech model Set generated is Suitable for use by a speech recognition System for recognizing spoken utterances for at least two different languages. The invention allows using a single Speech recognition unit with a single Speech model Set to perform Speech recognition on utterances from two or more languages. The method and apparatus make use of a group of a group of acoustic Sub-word units comprised of a first Subgroup of acoustic Sub-word units associated to a first language and a Second Subgroup of acoustic Sub-word units associated to a Second language where the first Subgroup and the Second Subgroup share at least one common acoustic Sub-word unit. The method and apparatus also make use of a plurality of letter to acoustic Sub-word unit rules Sets, each letter to acoustic Sub-word unit rules Set being associated to a different language. A Set of untrained speech models is trained on the basis of a training Set comprising speech tokens and their associated labels in combination with the group of acoustic Sub-word units and the plurality of letter to acoustic Sub-word unit rules Sets. The invention also provides a computer readable Storage medium comprising a program element for implementing the method for training a multilingual Speech model Set.” Sabourin at Abstract.

’925 Patent		
		“In accordance with another broad aspect, the invention provides a method for generating a multilingual speech model Set Suitable for use in a multilingual speech recognition System. The method comprises providing a group of acoustic Sub-word units having a first Subgroup of acoustic Sub-word units associated to a first language and a Second Subgroup of acoustic Sub-word units associated to a Second language. The first Subgroup and the Second Subgroup share at least one common acoustic Sub-word unit. The method further comprises providing a training Set comprising a plurality of entries, each entry having a speech token representative of a word and a label being an orthographic representation of the word. The method further comprises providing a Set of untrained speech models and training the Set of untrained speech models by utilizing the training Set, the plurality of letter to acoustic Sub-word unit rules Sets and the group of acoustic Sub-word units to derive the multilingual Speech model Set.” Sabourin at 2:57-3:7. “The invention provides a method for generating the multilingual speech model set shown in FIG.1. As shown in FIG. 2 of the drawings, the method comprises providing 200 a group of acoustic Sub-word units comprised of a first Subgroup of acoustic Sub-word units associated to a first language and a Second Subgroup of acoustic Sub-word units associated to a Second language. The first Subgroup and the Second Subgroup share at least one common acoustic Sub word unit. In a preferred embodiment, the group of acoustic Sub word unit is obtained by combining sub-word units from individual languages. For each language for which the multilingual Speech model Set is to be representative of, an inventory of sub-word units can be obtained. The acoustic Sub-word units are basic units of Sound in a language. These inventories may be produced by phoneticians or may be provided as off-the-shelf components. In a specific example, the Sub-word units are phonemes. The Skilled person in the art will readily see that other sub-word units may be used here without detracting from the spirit of the invention. The table below shows an example of a phoneme inventory for the Spanish language. …

’925 Patent		
		Typically each language possesses a unique inventory of acoustic Sub-word units. Preferably, each acoustic Sub-word unit in the acoustic Sub-word unit inventory for each language is associated to a feature description. The feature description provides a convenient means for establishing confusability rules for a given language and provide initialization information for the recognizer. The table below shows an example of feature descriptions of a Subset of the phonemes for the Spanish language. … The acoustic sub-word unit inventories for each language are then merged into a single group of acoustic sub-word units. Preferably acoustic sub-word unties that belong to the acoustic sub-word unit inventory of more than one language are stored only once in the group of acoustic sub-word units.” Sabourin at 4:25-5:14. “In its broad aspect, a nearest phoneme method includes initializing a speech model of a new phoneme on the basis of a speech model of a phoneme that is acoustically similar to the new phoneme. The phoneme acoustically similar to the new phoneme may be determined by a human which determines the closest sounding phoneme or by a heuristic algorithm implemented on a computer readable medium. In the preferred embodiment, the nearest phoneme is derived on the basis of a phonological similarity method. Alternatively, the speech models may be initialized by assigning random values to the initial models and by hand aligning the speech tokens with their corresponding transcriptions for a large number of speech tokens and training the initial model values. This alternative method is particularly useful when there are no speech models available. The invention provides a method for initializing a speech model set for a first language on the basis of a speech model set from a second language different from said first language. In a preferred embodiment, the invention makes use of the feature descriptions of the sub-word unit to generate initialization data for the speech models. In a specific example, suppose we have a first language for which speech models are available and a second language for which speech models are not available. In addition, suppose that in the second language, there is an acoustic sub-word unit,

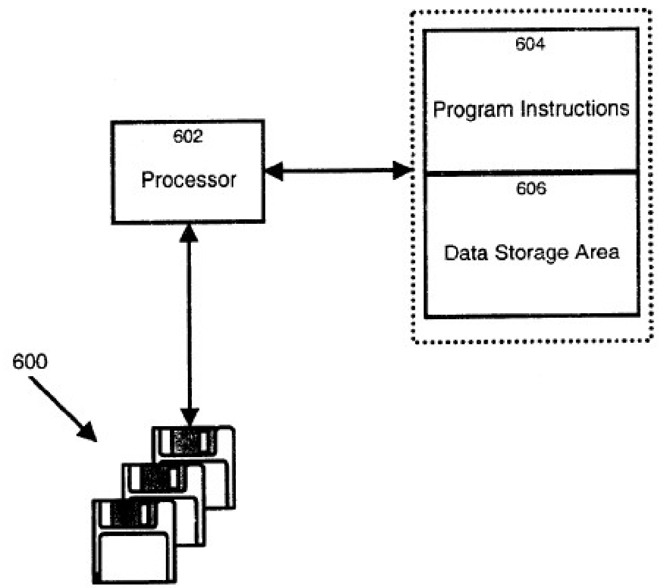
’925 Patent		
		herein designated as the new acoustic sub-word unit, that is not comprised in the acoustic sub-word inventory of the first language. The acoustic sub-word units common to the first language and the second language are initialized with the speech models associated to the first language. This invention provides a method for initializing the speech model of new acoustic sub-word unit on the basis of the known speech models associated to the first language more specifically by using the nearest phoneme as a basis to initialization. For example, say the nearest phoneme to /X/ according to a certain criteria is phoneme /Y/ and that the speech model for /Y/ is known. Initialization involves copying all the model parameters (eg. State transition weights) for model /Y/ into a model for /X/. This is particularly advantageous for initializing the speech model for sub-word units in a language for speech models are not available. The method will be described below for acoustic sub-word units being phonemes. The skilled person in the art will readily observe that this method may be applied to other types or acoustic sub-word units without detracting from the spirit of the invention.” Sabourin at 6:30-7:8. Sabourin at Figs. 2-5, 7. See e.g. Sabourin at claim 1 (and other claims).
1.b.2	“and wherein said re-estimating comprises partitioning said training data using said first decision network of said first speech recognizer.”	Sabourin specifically discloses partitioning training data using said first decision network. For example, Sabourin’s method includes associating the sub-word units in the speech models with the training set. *See, e.g.,* “As a variant, the training set 404 comprises a plurality of entries, the entries being representative of vocabulary items in two or more languages. The number of entries in the training set depends on the processing time available and the urgency of obtaining the results. The training set is first processed by the automatic transcription generator to extract the labels from each entry. The automatic transcription generator 500 is operative for processing the labels extracted from the training set 404 on the basis of a subset of the plurality of letter to acoustic sub-word unit rules sets 402 to derive a plurality of

’925 Patent		
		transcriptions. Each transcription in the plurality of transcriptions comprises a sequence of acoustic sub-word units of the group of acoustic sub-word units 400. A first transcription in the plurality of transcriptions corresponding to a certain vocabulary item is associated to a first language and a second transcription in the plurality of transcriptions corresponding to the same certain vocabulary item is associated to a second language different from the first language. The processing unit 408 further comprises a phoneme mapping unit 502 operatively connected to the first memory unit 400 and fourth memory unit 406 for associating the acoustic sub-word units in the group of acoustic sub-word units to respective speech models in said set of untrained speech models. The outputs of the automatic transcription generator 500 and of the phoneme mapping unit are operatively connected to a model training unit 504. The model training unit is also operatively coupled to the training set 404. The model training unit 504 is operative for processing the group of transcriptions received from the transcription generator 500 on the basis of a speech token of the corresponding entry in the training set 404 whereby training the set of untrained speech models 406 to derive the multilingual speech model set 410 on the basis of the method described in connection with FIG. 3.” Sabourin at 13:4-39. “The method also comprises providing 204 a training Set comprising a plurality of entries, each entry having a speech token representative of a Vocabulary item and a label being an orthographic representation of the Vocabulary item. The training Set may be obtained by Storing speech tokens and manually writing out each Vocabulary item associated to the Speech token or by having individuals read a predetermined Sequence of Vocabulary items. Such training Sets are also available as off-the-shelf components. In a specific example of implementation, the training Sets are reasonably accurate in that the Speech tokens have a high likelihood of corresponding to the written vocabulary items. A training Set is provided for each language that the multilingual speech model Set is to be representative of. In another embodiment, the labels can be assigned using a speech recognizer unit by Selecting the top scoring recognition candidate as the label for

’925 Patent		
		the utterance processed by the recognizer. The method also comprises providing 206 a set of untrained speech models. Each speech model is associated to an acoustic Sub-word unit in the group of acoustic Sub-word units. In a specific example, the untrained speech models are HMMS and are assigned a complex topology of about 80 means per phonemes and three HMMs states. The untrained Speech models may be obtained as off-the-shelf components or may be generated by using a nearest Sub word unit method. In the Specific example where the Sub word units are phonemes, a nearest phoneme method is used. The nearest phoneme method is described in detail below. The skilled person in the art will readily observe that it may be applied to acoustic Sub-word units other than phonemes without detracting from the Spirit of the invention. In its broad aspect, a nearest phoneme method includes initializing a speech model of a new phoneme on the basis of a speech model of a phoneme that is acoustically similar to the new phoneme. The phoneme acoustically Similar to the new phoneme may be determined by a human which determines the closest Sounding phoneme or by a heuristic algorithm implemented on a computer readable medium. In the preferred embodiment, the nearest phoneme is derived on the basis of a phonological Similarity method. Alternatively, the Speech models may be initialized by assigning random values to the initial models and by hand aligning the Speech tokens with their corresponding transcriptions for a large number of Speech tokens and training the initial model values. This alternative method is particularly useful when there are no speech models available.” Sabourin at 5:66-6:44 “The method also comprises training 208 the set of untrained Speech models by utilizing the training Set, the plurality of letter to acoustic Sub-word unit rules Sets and the group of acoustic Sub-word units to derive the multilingual speech model set 100. In a preferred embodiment, as shown in FIG. 3 of the drawings, training 208 the Set of untrained speech models comprises processing 300 at least Some of the entries in the training Set on the basis of a certain letter to acoustic Sub-word unit rules Set in the plurality of letter to acoustic Sub-word unit rules Sets to derive a group of transcriptions.

’925 Patent		
		In a Specific example, there is a plurality of training Sets, each Set being associated to a language. A given training Set associated to a given language is first processed to extract the labels from each entry. This operation will be readily available to those skilled in the art. The labels extracted from the given training Set are then processed by a given letter to acoustic Sub-word unit rules Set, the given letter to acoustic Sub-word unit rules Set being associated to the given language. The result of this processing is a group of transcriptions associated to a given language, the transcriptions in the group of transcriptions comprising a Sequence of acoustic Sub-word units of the group of acoustic Sub-word units. This processing is performed on all the training Sets of the plurality of training Sets. Optionally, the groups of transcriptions from each respective training Set are combined to form a compound training Set. As a variant, the training Set comprises a plurality of entries, the entries being representative of Vocabulary items in one or more languages. The training Set is first processed to extract the labels from each entry. The labels extracted from the training Set are then processed by a Subset of the plurality of letter to acoustic Sub-word unit rules Sets to derive a plurality of transcriptions. A first transcription in the plurality of transcriptions corresponding to a certain vocabulary item is associated to a first language and a Second transcription in the plurality of transcriptions corresponding to the same certain vocabulary item is associated to a Second language different from the first language. As yet another variant, prosodic rules may be used to process the transcriptions generated by with the letter to acoustic unit rules Sets to derive Surface form transcriptions. The prosodic rules are associated to respective languages may be applied to the groups of transcriptions associated to the same languages. Alternatively, a Subset of the prosodic rules may be applies to the transcriptions associated to languages different from the languages associated to the prosodic rules being considered. In accordance with a preferred embodiment, training the Set of untrained Speech models further comprises associating 302 the acoustic Sub-word units in the group of acoustic Sub-word units to respective speech models in the Set of untrained speech models. The training of the Set of untrained speech models further comprises processing 304 the group of transcriptions generated at step 300 on the basis of a speech token of the corresponding entry in the training Set whereby training the Set of untrained speech

’925 Patent		
		models to derive the multilingual Speech model Set. In a specific example of implementation, the Speech tokens in Said training Set are processed by a feature extraction unit to convert the Speech Signals into a Set of numeric values for Storage in a vector representative of acoustic information. A specific example of a vector is a cepstral vector. Each speech token is associated to a set of cepstral vectors, one vector being associated to a frame of the Speech token. The cepstral vectors are then aligned with the phonemes of the transcription corresponding to the Speech token. In a specific example, for the initial alignment, the cepstral vectors of the Speech token are randomly divided between the phonemes of the transcription. In another specific example, vector quantization or Kohonen associative mapping is used to provide the initial alignment of the cepstral vectors and the phonemes. These cepstral vectors are then used to condition the untrained speech models thereby training the Speech models. In a preferred embodiment, the Speech models are trained using a maximum likelihood method. The speech models are HMMs having a set of States and transitions between these States, each State represented by a State probability and a set of transition probabilities. The state probability is modeled as a mixture of Gaussian distributions where each Gaussian distribution has a means and a covariance matrix. The means and covariance matrices may be shared be other models without detracting from the spirit of the invention. The transition probabilities are exponential distributions with mean u. In a Specific example, the untrained speech models are trained using maximum a posteriori adaptation (MAP) as described in Gauvain J.-L. and Lee C. -H. (1994) “maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains' IEEE Trans. Speech Audio Process. 2, pp. 291-298. The content of this document is hereby incorporated by reference. Other methods of training Speech models on the basis of Speech tokens may be used here without detracting from the Spirit of the invention. In a preferred embodiment Several training iterations are performed to condition the Speech models. The iterations are performed by obtaining another assignation of the vectors to the phonemic transcription of the Speech label and restarting the training.” Sabourin at 8:49-10:23.

’925 Patent		
		Sabourin at Figs. 2-5, 7. See e.g. Sabourin at claim 1 (and other claims).
Claim 14		
14.pre	A machine-readable storage medium, having stored thereon a computer program having a plurality of code sections executable by a machine for causing the machine to automatically generate from a first speech recognizer a second speech recognizer, said first speech recognizer comprising a first acoustic model with a first decision network and corresponding first phonetic contexts, and said second speech recognizer being adapted to a specific domain, said machine-readable storage causing the machine to perform the steps of:	*See claim 1, including 1.pre.a-c.* Sabourin further discloses that a machine readable storage with a computer program having code that executes the described limitations. *See e.g.,* *See* Sabourin at Fig. 6.  “The invention relates to a method and apparatus for training a multilingual speech model Set. The multilingual Speech model Set generated is Suitable for use by a speech recognition System for recognizing spoken utterances for at least two different

’925 Patent		
		languages. The invention allows using a single Speech recognition unit with a single Speech model Set to perform Speech recognition on utterances from two or more languages. The method and apparatus make use of a group of a group of acoustic Sub-word units comprised of a first Subgroup of acoustic Sub-word units associated to a first language and a Second Subgroup of acoustic Sub-word units associated to a Second language where the first Subgroup and the Second Subgroup share at least one common acoustic Sub-word unit. The method and apparatus also make use of a plurality of letter to acoustic Sub-word unit rules Sets, each letter to acoustic Sub-word unit rules Set being associated to a different language. A Set of untrained speech models is trained on the basis of a training Set comprising speech tokens and their associated labels in combination with the group of acoustic Sub-word units and the plurality of letter to acoustic Sub-word unit rules Sets. The invention also provides a computer readable Storage medium comprising a program element for implementing the method for training a multilingual Speech model Set.” Sabourin at Abstract. “In accordance with another broad aspect, the invention provides an apparatus for generating a multilingual speech model set. In accordance with another broad aspect, the invention provides a computer readable storage medium containing a program element suitable for use on a computer having a memory, the program element being suitable for generating a multilingual speech model set.” Sabourin at 3:8-15. “In a specific embodiment, the apparatus depicted in FIGS. 4 and 5 comprises a processor coupled to a computer reasonable storage medium, the computer readable storage medium comprising a program element for execution by the processor for implementing the processing unit 408. …

’925 Patent		
		The above-described method of generating a multilingual speech model set can also be implemented on any suitable computing platform as show in FIG 6.” Sabourin at 13:40-50. “10. A computer readable storage medium comprising a data structure containing a multilingual speech model set generated by the method defined in claim 1.” Sabourin at 15:65-67. Sabourin claim 11-24.
14.a	based on said first acoustic model, generating a second acoustic model with a second decision network and corresponding second phonetic contexts for said second speech recognizer by re-estimating said first decision network and said corresponding first phonetic contexts based on domain-specific training data, wherein said first decision network and said second decision network utilize a phonetic decision tree to perform speech recognition operations,	*See claim 1, including 1.a.1-2, and 14.pre.*
14.b	wherein the number of nodes in the second decision network is not fixed by the number of nodes in the first decision network, and wherein said re-estimating	*See claim 1, including 1.b.1-2, and 14.pre, 14.a.*

’925 Patent		
	comprises partitioning said training data using said first decision network of said first speech recognizer.	
Claim 27		
27.pre	“A computerized method of generating a second speech recognizer comprising the steps of:	*See claim 1, including 1.pre.a-c.*
27.a	identifying a first speech recognizer of a first domain comprising a first acoustic model with a first decision network and corresponding first phonetic contexts;”	*See claim 1, including 1.pre.b and 27.pre.*
27.b	“receiving domain-specific training data of a second domain; and”	Sabourin discloses receiving domain-specific training data of a second domain. *See claim 1 including 1.pre.c and 1.a.1. See also* “A training set is provided for each language that the multilingual speech model set is to be representative of.” Sabourin at 6:10-12. “In a specific example, there is a plurality of training sets, each set being associated to a language.” Sabourin at 8:60-61. Sabourin at Fig. 3; *see also* Figs. 2, 4, 5. *See, e.g.*, Sabourin, Claim 1 (and other claims).

’925 Patent		
27.c	“based on the first speech recognizer and the domain-specific training data, generating a second acoustic model of said first domain and said second domain comprising a second acoustic model with a second decision network and corresponding second phonetic contexts, wherein the first domain comprises at least a first language, wherein the second domain comprises at least a second language, and wherein the second speech recognizer is a multi-lingual speech recognizer.”	Sabourin discloses based on the first speech recognizer and the domain-specific training data, generating a second acoustic model of said first domain and said second domain comprising a second acoustic model with a second decision network and corresponding second phonetic contexts. *See claim 1 including 1.pre.c and 1.a.1.* Further, Sabourin discloses a multilingual system where there are two different language domains as recited. “The invention relates to a method and apparatus for training a multilingual speech model Set. The multilingual Speech model Set generated is Suitable for use by a speech recognition System for recognizing spoken utterances for at least two different languages. The invention allows using a single Speech recognition unit with a single Speech model Set to perform Speech recognition on utterances from two or more languages. The method and apparatus make use of a group of a group of acoustic Sub-word units comprised of a first Subgroup of acoustic Sub-word units associated to a first language and a Second Subgroup of acoustic Sub-word units associated to a Second language where the first Subgroup and the Second Subgroup share at least one common acoustic Sub-word unit. The method and apparatus also make use of a plurality of letter to acoustic Sub-word unit rules Sets, each letter to acoustic Sub-word unit rules Set being associated to a different language. A Set of untrained speech models is trained on the basis of a training Set comprising speech tokens and their associated labels in combination with the group of acoustic Sub-word units and the plurality of letter to acoustic Sub-word unit rules Sets. The invention also provides a computer readable Storage medium comprising a program element for implementing the method for training a multilingual Speech model Set.” Sabourin at Abstract. “A deficiency of the above-described method is that the Speech recognition System requires as an input the language associated to the input utterance, which may not be readily available to the Speech recognition System. Usually, obtaining the language requires prompting the user for the language of use thereby requiring an additionally

’925 Patent		
		Step in the Service being provided by the Speech recognition enabled system which may lower the level of satisfaction of the user with the system as a whole. Another deficiency of the above noted method is the costs associated to developing and maintaining a Speech model Set for each language the Speech recognition System is adapted to recognize. More Specifically, each speech model Set must be trained individually, a task requiring manpower for each individual language thereby increasing significantly the cost of Speech recognition Systems operating in multilingual environments with respect to Systems operating in unilingual environments. In addition, the above-described method requires the Storage of a speech model Set for each language in memory thereby increasing the cost of the Speech recognition System in terms of memory requirements. Finally, the above described method requires testing a speech model Set for each language thereby increasing the testing cost of the Speech recognition System for each language the Speech recognition System is adapted to recognize. Thus, there exists a need in the industry to refine the process of training Speech models So as to obtain an improved multilingual Speech model Set capable of being used by a speech recognition System for recognizing spoken utterances for at least two different languages.” Sabourin at 1:55-2:17. “In accordance with another broad aspect, the invention provides a method for generating a multilingual speech model Set Suitable for use in a multilingual speech recognition System. The method comprises providing a group of acoustic Sub-word units having a first Subgroup of acoustic Sub-word units associated to a first language and a Second Subgroup of acoustic Sub-word units associated to a Second language. The first Subgroup and the Second Subgroup share at least one common acoustic Sub-word unit. The method further comprises providing a training Set comprising a plurality of entries, each entry having a speech token representative of a word and a label being an orthographic representation of the word. The method further comprises providing a Set of untrained speech models and training the Set of untrained speech models by utilizing the training Set, the plurality of letter to acoustic Sub-word unit

’925 Patent		
		rules Sets and the group of acoustic Sub-word units to derive the multilingual Speech model Set.” Sabourin at 2:57-3:7. “The invention provides a method for generating the multilingual speech model set shown in FIG.1. As shown in FIG. 2 of the drawings, the method comprises providing 200 a group of acoustic Sub-word units comprised of a first Subgroup of acoustic Sub-word units associated to a first language and a Second Subgroup of acoustic Sub-word units associated to a Second language. The first Subgroup and the Second Subgroup share at least one common acoustic Sub word unit. In a preferred embodiment, the group of acoustic Sub word unit is obtained by combining sub-word units from individual languages. For each language for which the multilingual Speech model Set is to be representative of, an inventory of sub-word units can be obtained. The acoustic Sub-word units are basic units of Sound in a language. These inventories may be produced by phoneticians or may be provided as off-the-shelf components. In a specific example, the Sub-word units are phonemes. The Skilled person in the art will readily see that other sub-word units may be used here without detracting from the spirit of the invention.” Sabourin at 4:25-45. “The training of the set of untrained speech models further comprises processing 304 the group of transcriptions generated at step 300 on the basis of a speech token of the corresponding entry in the training Set whereby training the Set of untrained speech models to derive the multilingual Speech model Set.” Sabourin at 9:34-39. *See* Sabourin at Figs. 2-7. *See, e.g.*, Sabourin, Claim 1 (and other claims).

APPENDIX A-3
Invalidity Claim Chart for U.S. Pat. No. 6,999,925 ('925 patent)
U.S. Pat. No. 6,324,510 ("Waibel")[1]

On October 16, 2020, Nuance narrowed the asserted '925 patent claims to 1, 14 and 27. Nuance's Disclosure Limiting Asserted Claims of Phase 1 Patents. Claim 27 of the '925 patent is anticipated and/or rendered obvious by Waibel alone or in combination with at least the following references. The chart below discloses how prior art references identified by Omilia disclose, either expressly or inherently, and/or render obvious these asserted claims:

(1) U.S. Pat. No. 6,912,499, Sabourin et al., filed August 31, 1999 ("Sabourin")[2]

This invalidity claim chart is based on Omilia's present understanding of the asserted claims, the Court's August 6, 2020 Claim Construction Order (ECF. No. 157), Nuance's initial and all subsequent supplemental Infringement Contentions, its July 7, 2020 Response to Omilia's Supplemental Non-Infringement and Invalidity Responses, Nuance's Response to Omilia's Interrogatory No. 9, and Omilia's investigation to date. Omilia is not adopting Nuance's apparent constructions, nor is Omilia admitting to the accuracy of any particular construction. Omilia reserves all rights to amend this invalidity claim chart if Nuance further amends its Infringement Contentions, or in response to any statements made by Nuance at any phase of this litigation. In addition, Discovery is still ongoing, and Omilia reserves its right to supplement its contentions with additional evidence as discovery continues.

The citations provided are exemplary and do not necessarily include each and every disclosure of the limitation in the references. Omilia has cited the most relevant portions of the identified prior art. Other portions of the identified prior art may additionally disclose, either expressly or inherently, and/or render obvious one or more limitations of the asserted claims. To provide context or aid in understanding the prior art and the state of the art at the time of the alleged invention, Omilia reserves the right to rely on: (1) uncited portions of the identified prior art; (2) other prior art not identified herein; (3) references that show the state of the art (irrespective of whether such references themselves qualify as prior art to the asserted patents); (4) factual testimony from the inventors or authors of the prior art references, or purveyors of prior art devices; and/or (5) expert testimony.

Citations to a particular drawing or figure in the accompanying charts should be understood to include the description of the drawing or figure, as well as any text associated with the drawing or figure. Citations to particular text concerning a drawing or figure, should also be understood to encompasses that drawing or figure as well. The identified prior art expressly and/or inherently discloses the features of the asserted claims under all proposed constructions of the asserted claims. To establish the inherency of certain features of the prior

[1] Waibel was filed on November 6, 1998 and is prior art at least under 35 U.S.C. § 102(a) & 102(e).

[2] Sabourin was filed on August 31, 1999 and is prior art at least under 35 U.S.C. § 102(a) & 102(e).

art to invalidate the asserted claims, Omilia may rely on cited or uncited portions of the prior art, other documents, factual testimony, and expert testimony. To the extent that the Court finds that this reference does not explicitly teach certain limitations in the asserted claims, such limitations would have been inherent and/or obvious and Omilia reserves the right to supplement or amend this claim chart to address such a finding.

<u>’925 Patent</u>		
Claim 27		
27.pre	“A computerized method of generating a second speech recognizer comprising the steps of:	Waibel discloses a computerized method of generating a second speech recognizer. *See e.g.*, “A method of organizing an acoustic model for speech recognition is comprised of the steps of calculating a measure of acoustic dissimilarity of subphonetic units. A clustering technique is recursively applied to the subphonetic units based on the calculated measure of acoustic dissimilarity to automatically generate a hierarchically arranged model. Each application of the clustering technique produces another level of the hierarchy with the levels progressing from the least specific to the most specific. A technique for adapting the structure and size of a trained acoustic model to an unseen domain using only a small amount of adaptation data is also disclosed.” Waibel, at Abstract. “The present invention is also directed to a method of structurally adapting a hierarchical acoustic model having both nodes and leaves to a new domain. The method is comprised of the steps of identifying nodes that receive more than a predetermined amount of adaptation data and adapting the local estimators of conditional posteriors and priors of the identified nodes using data from the new domain. A user-specified quantity of the non-identified nodes are removed and leaves are created, where needed, to replace the removed nodes. All of the HMM states are related to the new leaves such that they share a single model represented by the new leaves.” Waibel, at 2:66–3:10.

’925 Patent		
		“The present invention is directed to a method of organizing an acoustic model for speech recognition comprised of the steps of calculating a measure of acoustic dissimilarity of subphonetic units. Recursively clustering the subphonetic units based on the calculated measure automatically generates a hierarchically arranged model. An apparatus for performing the method is also disclosed.” Waibel, at 2:52–58. “Starting from an initial set of decision tree clustered, context-dependent, subphonetic units, the present invention uses an aglommerative clustering algorithm across monophones to automatically design a tree-structured decomposition of posterior probabilities which is instantiated with thousands of small neural network estimators at each of the nodes of the tree.” Waibel, at 2:59–65. “FIG. 1 illustrates a computer on which the present invention may be practiced.” Waibel, at 3:41-42; *see also* Fig. 1. “The methods of the present invention may be carried out on a computer 10 of the type illustrated in FIG. 1. It is anticipated that the methods of the present invention will be embodied in software and conventionally stored such as on the computer’s hard drive 12, a floppy dish 14, or other storage medium. When the computer 10 executes software which embodies the methods of the present invention, the computer 10 becomes the means necessary for performing the various steps of the method.” Waibel, at 3:56-64. Waibel, Claims 16-36, in particular claim 23.
27.a	identifying a first speech recognizer of a first domain	Waibel discloses a first speech recognizer with a first acoustic model that includes a decision network and corresponds to phonetic contexts at least because Waibel is describes

’925 Patent		
	comprising a first acoustic model with a first decision network and corresponding first phonetic contexts;”	HMMs and HMM-based speech recognition systems which make use of decision networks with corresponding phonetic context as described in Waibel. *See e.g.*, “Another problem with current HMM-based speech recognition technology is that it suffers from domain dependence. Over the years, the community has validated and commercialized the technology based on standardized training and test sets in restricted domains, such as the Wall Street Journal (WSJ) (business newspaper texts), Switchboard (SWB) (spontaneous telephone conversations) and Broadcast News (BN) (radio/tv news shows). Performance of systems trained on such domains typically drops significantly when applied to a different domain, especially with changing speaking style, e.g. when moving from read speech to spontaneous speech. D. L. Thomson, “Ten Case Studies of the Effect of Field Conditions on Speech Recognition Errors”, *Proceedings of the IEEE ASRU Workshop*, Santa Barbara, 1997. For instance, performance of a recognizer trained on WSJ typically decreases severely when decoding SWB data. Several factors can be held responsible for the strong domain dependence of current statistical speech recognition systems. One is constrained quality, type or recording conditions of domain specific speech data (read, conversational, spontaneous speech/noisy, clean recordings/presence of acoustic background sources, etc.). Another is vocabulary and language model dependence of phonetic context modeling based on phonetic decision trees. That implies a strong dependence of allophonic models on the specific domain. Another factor is domain dependent optimization of size of acoustic model based on amount of available training data and/or size of vocabulary. While the first of the above-mentioned factors is typically addressed by some sort of speaker and/or environment adaptation technique, the latter two factors are usually not adequately addressed in cross-domain applications.” Waibel, at 1:64-2:28. “The present invention is also directed to a method of structurally adapting a hierarchical acoustic model having both nodes and leaves to a new domain. The method is comprised of the steps of identifying nodes that receive more than a predetermined amount of adaptation data and adapting the local estimators of conditional posteriors and priors of the identified nodes using data from the new domain. A user-specified quantity of the non-

’925 Patent		
		identified nodes are removed and leaves are created, where needed, to replace the removed nodes. All of the HMM states are related to the new leaves such that they share a single model represented by the new leaves.” Waibel, at 2:66–3:10. “The disclosed method allows effective adaptation of the structure and size of a trained acoustic model to an unseen domain using only a small amount of adaptation data. The present invention benefits from the multi-level, hierarchical representation of the context-dependent acoustic model. In contrast to approaches based on acoustic adaptation only, the present invention uses an estimate of the a-priori distribution of modeled HMM states on the new domain to dynamically downsize or prune the tree-structured acoustic model. In that manner, the present invention accounts for differences in vocabulary size and adjusts to the specificity of phonetic context observed in the new domain.” Waibel, at 3:11-22. “By adapting the specificity of the acoustic model, improved performance can be obtained with very little requirements for adaptation data. Furthermore, the present invention compensates over fitting effects particularly when targeting a domain with a much smaller vocabulary. The present invention may also be applied to downsize/prune an acoustic model to any desired size to accommodate computing and/or memory resource limitations. Those, and other advantages and benefits of the present invention, will become apparent from reading the Description Of The Preferred Embodiment hereinbelow.” Waibel, at 3:23-33. “An interesting property of HNNs that can be exploited for structural adaptation is that partially computed posterior probabilities at all crossed paths in every horizontal cross section of the tree constitute a legal posterior probability distribution over a reduced (merged) set of leaves. A starting point for structural adaptation is an HNN constructed and trained on a domain exhibiting sufficiently rich diversity in phonetic context to provide a

'925 Patent		
		basis for any new, unseen domain. To adapt this baseline for any new, smaller domain typically exhibiting very different specificity of phonetic context, the following steps are performed: 1. Take the baseline HNN tree (circles=nodes, squares=leaves) (FIG. 5A) 2. Select nodes that receive more than a predetermined, sufficiently large amount of adaptation data (mincount) and adapt their local estimators of conditional posteriors and priors using adaptation data from the new domain. (FIG. 5B) 3. Remove all nodes that receive less than a predetermined amount of adaptation data. Create new leaf nodes (squares) in place of the root nodes of pruned subtrees. (FIG. 5C) 4. Finally, merge leaf nodes of pruned subtrees. (FIG.5D) Tie all HMM states corresponding to the leaves of pruned subtrees in the original tree such that they share a single model, represented by the newly created leaves." Waibel, at 6:29-6:55. "A method of organizing an acoustic model for speech recognition is comprised of the steps of calculating a measure of acoustic dissimilarity of subphonetic units. A clustering technique is recursively applied to the subphonetic units based on the calculated measure of acoustic dissimilarity to automatically generate a hierarchically arranged model. Each application of the clustering technique produces another level of the hierarchy with the levels progressing from the least specific to the most specific. A technique for adapting the structure and size of a trained acoustic model to an unseen domain using only a small amount of adaptation data is also disclosed." Waibel, at Abstract. "The present invention is directed to a method of organizing an acoustic model for speech recognition comprised of the steps of calculating a measure of acoustic dissimilarity of

’925 Patent		
		subphonetic units. Recursively clustering the subphonetic units based on the calculated measure automatically generates a hierarchically arranged model. An apparatus for performing the method is also disclosed.” Waibel, at 2:52–58. “Starting from an initial set of decision tree clustered, context-dependent, subphonetic units, the present invention uses an aglommerative clustering algorithm across monophones to automatically design a tree-structured decomposition of posterior probabilities which is instantiated with thousands of small neural network estimators at each of the nodes of the tree.” Waibel, at 2:59–65. “In contrast to conventional mixtures of Gaussians based acoustic models, the HNN framework of the present invention does not require additional structures to reduce the complexity of model evaluation. The tree structure itself can be exploited to control the speed-accuracy trade-off. The size of the tree, and hence the degree of accuracy, may be dynamically adapted based on the requirements and data available for a given task. The evaluation of posterior state probabilities follows a path from root node to a specific leaf in the HNN, multiplying all estimates of conditional posteriors along the way. Subtrees can be pruned by closing paths whenever the partial probability falls below a suitable threshold. This can be performed dynamically during speech recognition. This way the evaluation of a significant amount of networks at the bottom of the HNN can be avoided, possibly at the cost of increased error rate.” Waibel, at 7:9–24. *See also* Waibel, Claims 16-36.

’925 Patent		
27.b	“receiving domain-specific training data of a second domain; and”	Waibel discloses receiving domain-specific training data of a second domain. Waibel describes using “adaptation data” to adapt the decision tree model. *See e.g.*, “The present invention is also directed to a method of structurally adapting a hierarchical acoustic model having both nodes and leaves to a new domain. The method is comprised of the steps of identifying nodes that receive more than a predetermined amount of adaptation data and adapting the local estimators of conditional posteriors and priors of the identified nodes using data from the new domain. A user-specified quantity of the non-identified nodes are removed and leaves are created, where needed, to replace the removed nodes. All of the HMM states are related to the new leaves such that they share a single model represented by the new leaves.” Waibel, at 2:66–3:10. “The disclosed method allows effective adaptation of the structure and size of a trained acoustic model to an unseen domain using only a small amount of adaptation data. The present invention benefits from the multi-level, hierarchical representation of the context-dependent acoustic model. In contrast to approaches based on acoustic adaptation only, the present invention uses an estimate of the a-priori distribution of modeled HMM states on the new domain to dynamically downsize or prune the tree-structured acoustic model. In that manner, the present invention accounts for differences in vocabulary size and adjusts to the specificity of phonetic context observed in the new domain.” Waibel, at 3:11–22.
27.c	“based on the first speech recognizer and the domain-specific training data, generating a second acoustic model of said first domain and said second domain comprising a second	Waibel discloses based on the first speech recognizer and the domain-specific training data, generating a second acoustic model of said first domain and said second domain comprising a second acoustic model with a second decision network and corresponding second phonetic contexts. In Waibel, the generating of a new speech recognizer includes re-estimating said first decision network and said corresponding first phonetic contexts based on domain-

’925 Patent		
	acoustic model with a second decision network and corresponding second phonetic contexts, wherein the first domain comprises at least a first language, wherein the second domain comprises at least a second language, and wherein the second speech recognizer is a multi-lingual speech recognizer.”	specific training data, at least because the method adapts the structure and size of the acoustic model and its phonetic contexts. *See e.g.*, “A method of organizing an acoustic model for speech recognition is comprised of the steps of calculating a measure of acoustic dissimilarity of subphonetic units. A clustering technique is recursively applied to the subphonetic units based on the calculated measure of acoustic dissimilarity to automatically generate a hierarchically arranged model. Each application of the clustering technique produces another level of the hierarchy with the levels progressing from the least specific to the most specific. A technique for adapting the structure and size of a trained acoustic model to an unseen domain using only a small amount of adaptation data is also disclosed.” Waibel, at Abstract. “The present invention is also directed to a method of structurally adapting a hierarchical acoustic model having both nodes and leaves to a new domain. The method is comprised of the steps of identifying nodes that receive more than a predetermined amount of adaptation data and adapting the local estimators of conditional posteriors and priors of the identified nodes using data from the new domain. A user-specified quantity of the non-identified nodes are removed and leaves are created, where needed, to replace the removed nodes. All of the HMM states are related to the new leaves such that they share a single model represented by the new leaves.” Waibel, at 2:66–3:10. “The disclosed method allows effective adaptation of the structure and size of a trained acoustic model to an unseen domain using only a small amount of adaptation data. The present invention benefits from the multi-level, hierarchical representation of the context-dependent acoustic model. In contrast to approaches based on acoustic adaptation only, the present invention uses an estimate of the a-priori distribution of modeled HMM states on the new domain to dynamically downsize or prune the tree-structured acoustic model. In

’925 Patent		
		that manner, the present invention accounts for differences in vocabulary size and adjusts to the specificity of phonetic context observed in the new domain.” Waibel, at 3:11–22. “By adapting the specificity of the acoustic model, improved performance can be obtained with very little requirements for adaptation data. Furthermore, the present invention compensates over fitting effects particularly when targeting a domain with a much smaller vocabulary. The present invention may also be applied to downsize/prune an acoustic model to any desired size to accommodate computing and/or memory resource limitations. Those, and other advantages and benefits of the present invention, will become apparent from reading the Description Of The Preferred Embodiment hereinbelow.” Waibel, at 3:23–33. “An interesting property of HNNs that can be exploited for structural adaptation is that partially computed posterior probabilities at all crossed paths in every horizontal cross section of the tree constitute a legal posterior probability distribution over a reduced (merged) set of leaves. A starting point for structural adaptation is an HNN constructed and trained on a domain exhibiting sufficiently rich diversity in phonetic context to provide a basis for any new, unseen domain. To adapt this baseline for any new, smaller domain typically exhibiting very different specificity of phonetic context, the following steps are performed: 1. Take the baseline HNN tree (circles=nodes, squares=leaves) (FIG. 5A) 2. Select nodes that receive more than a predetermined, sufficiently large amount of adaptation data (mincount) and adapt their local estimators of conditional posteriors and priors using adaptation data from the new domain. (FIG. 5B)

’925 Patent		
		3. Remove all nodes that receive less than a predetermined amount of adaptation data. Create new leaf nodes (squares) in place of the root nodes of pruned subtrees. (FIG. 5C) 4. Finally, merge leaf nodes of pruned subtrees. (FIG.5D) Tie all HMM states corresponding to the leaves of pruned subtrees in the original tree such that they share a single model, represented by the newly created leaves.” Waibel, at 6:28–55. “In contrast to conventional mixtures of Gaussians based acoustic models, the HNN framework of the present invention does not require additional structures to reduce the complexity of model evaluation. The tree structure itself can be exploited to control the speed-accuracy trade-off. The size of the tree, and hence the degree of accuracy, may be dynamically adapted based on the requirements and data available for a given task. The evaluation of posterior state probabilities follows a path from root node to a specific leaf in the HNN, multiplying all estimates of conditional posteriors along the way. Subtrees can be pruned by closing paths whenever the partial probability falls below a suitable threshold. This can be performed dynamically during speech recognition. This way the evaluation of a significant amount of networks at the bottom of the HNN can be avoided, possibly at the cost of increased error rate.” Waibel, at 7:9–24. “The present invention maintains the advantages of discriminative training while circumventing the limitations of standard connectionist acoustic models. Furthermore, HNN acoustic models incorporate the Structure for Speaker adaptation and Scoring Speed-up algorithms that usually require additional effort in traditional mixture densities acoustic models. The present invention enables effective adaptation of the Structure of a tree-structured hierarchical connectionist acoustic model to unseen new domains. In contrast to existing architectures and adaptation techniques, the present invention not only compensates for mismatches in acoustic Space, but adapts to differing Specificity of phonetic context in

’925 Patent		
		unseen domains by adapting node priors and pruning defective parts of the modeling hierarchy.” Waibel, at 7:45-58: *See also* Waibel, claims 16-36 While Waibel does not explicitly contemplate that the final speech recognizer will be able to understand multiple languages, Waibel contemplates that its solution avoids the cost with adapting acoustic models for example, in connection with different applications within a language. *See* Waibel at 2:66-3:33 and 2:29-42. A POSITA would understand that this same adaptation could be performed with a pre-existing multi-lingual recognizer or used to create a multilingual recognizer. This was a common application and there was a need for modified recognizers. For example, multilingual speech recognizers were well known at the time of the ’925 patent. *See, e.g.*, Schultz et al., “*Polyphone Decision Tree Specialization for Language Adaptation*”, ICASSP-2000, Istanbul, Turkey, Jun. 2000 (Section 2.2. describing generating and use of multilingual speech recognizers). A POSITA would have known that multilingual speech recognizers were well known and desirable. *See* Exhibit A-1, (Sabourin Chart, claim 27.c). Given this disclosure, the creation of a multilingual second recognizer was obvious in light of Waibel alone, or in combination with Sabourin. The motivation to combine these references would at least include: • Combining prior art elements according to known methods to yield predictable results (multilingual recognizers were known using similar techniques for predictable results); • Simple substitution of one known element for another to obtain predictable results (substituting the domain from just one language to two);

’925 Patent		
		• Use of known technique to improve similar devices (methods, or products) in the same way (same technique of expanding domain of the recognizer to include multiple languages was known)); • Applying a known technique to a known device (method, or product) ready for improvement to yield predictable results (multilingual recognizers were known using similar known techniques for predictable results); • “Obvious to try”—choosing from a finite number of identified, predictable solutions, with a reasonable expectation of success (obvious to try to recognize more than one language); • Market forces and benefits associated with the known benefits of multilingual recognizers were predictable to POSA • Teaching of prior art would have lead a POSA to combine the references to arrive at a multilingual recognizer. For example, Sabourin explicitly discloses a multilingual system. Sabourin also discloses that there are two different language domains as recited. “The invention relates to a method and apparatus for training a multilingual speech model Set. The multilingual Speech model Set generated is Suitable for use by a speech recognition System for recognizing spoken utterances for at least two different languages. The invention allows using a single Speech recognition unit with a single Speech model Set to perform Speech recognition on utterances from two or more languages. The method and apparatus make use of a group of a group of acoustic Sub-word units comprised of a first Subgroup of acoustic Sub-word units associated to a first language and a Second Subgroup of acoustic Sub-word units associated to a Second language where the first Subgroup and the Second Subgroup share at least one common acoustic Sub-word unit. The method and apparatus also make use of a plurality of letter to acoustic Sub-word unit rules Sets, each letter to acoustic Sub-word unit rules Set being associated to a different language. A Set of untrained speech models is trained on the basis of a training Set

’925 Patent		
		comprising speech tokens and their associated labels in combination with the group of acoustic Sub-word units and the plurality of letter to acoustic Sub-word unit rules Sets. The invention also provides a computer readable Storage medium comprising a program element for implementing the method for training a multilingual Speech model Set.” Sabourin at Abstract. “A deficiency of the above-described method is that the Speech recognition System requires as an input the language associated to the input utterance, which may not be readily available to the Speech recognition System. Usually, obtaining the language requires prompting the user for the language of use thereby requiring an additionally Step in the Service being provided by the Speech recognition enabled system which may lower the level of satisfaction of the user with the system as a whole. Another deficiency of the above noted method is the costs associated to developing and maintaining a Speech model Set for each language the Speech recognition System is adapted to recognize. More Specifically, each speech model Set must be trained individually, a task requiring manpower for each individual language thereby increasing significantly the cost of Speech recognition Systems operating in multilingual environments with respect to Systems operating in unilingual environments. In addition, the above-described method requires the Storage of a speech model Set for each language in memory thereby increasing the cost of the Speech recognition System in terms of memory requirements. Finally, the above described method requires testing a speech model Set for each language thereby increasing the testing cost of the Speech recognition System for each language the Speech recognition System is adapted to recognize. Thus, there exists a need in the industry to refine the process of training Speech models So as to obtain an improved multilingual Speech model Set capable of being used by a speech recognition System for recognizing spoken utterances for at least two different languages.” Sabourin at 1:55-2:17. “In accordance with another broad aspect, the invention provides a method for generating a multilingual speech model Set Suitable for use in a multilingual speech recognition

’925 Patent		
		System. The method comprises providing a group of acoustic Sub-word units having a first Subgroup of acoustic Sub-word units associated to a first language and a Second Subgroup of acoustic Sub-word units associated to a Second language. The first Subgroup and the Second Subgroup share at least one common acoustic Sub-word unit. The method further comprises providing a training Set comprising a plurality of entries, each entry having a speech token representative of a word and a label being an orthographic representation of the word. The method further comprises providing a Set of untrained speech models and training the Set of untrained speech models by utilizing the training Set, the plurality of letter to acoustic Sub-word unit rules Sets and the group of acoustic Sub-word units to derive the multilingual Speech model Set.” Sabourin at 2:57-3:7. “The invention provides a method for generating the multilingual speech model set shown in FIG.1. As shown in FIG. 2 of the drawings, the method comprises providing 200 a group of acoustic Sub-word units comprised of a first Subgroup of acoustic Sub-word units associated to a first language and a Second Subgroup of acoustic Sub-word units associated to a Second language. The first Subgroup and the Second Subgroup share at least one common acoustic Sub word unit. In a preferred embodiment, the group of acoustic Sub word unit is obtained by combining sub-word units from individual languages. For each language for which the multilingual Speech model Set is to be representative of, an inventory of sub-word units can be obtained. The acoustic Sub-word units are basic units of Sound in a language. These inventories may be produced by phoneticians or may be provided as off-the-shelf components. In a specific example, the Sub-word units are phonemes. The Skilled person in the art will readily see that other sub-word units may be used here without detracting from the spirit of the invention.” Sabourin at 4:25-45. “The training of the set of untrained speech models further comprises processing 304 the group of transcriptions generated at step 300 on the basis of a speech token of the

’925 Patent		
		corresponding entry in the training Set whereby training the Set of untrained speech models to derive the multilingual Speech model Set.” Sabourin at 9:34-39. *See* Sabourin at Figs. 1-7. It would be obvious to a person having ordinary skill in the art to combine Waibel with Sabourin. Both Waibel and Sabourin are in the same field of art. A person having ordinary skill in the art would be motivated to combine Waibel with Sabourin. A person having ordinary skill in the art considering Waibel’s disclosure that there is a need for a model that “is easily integrated into existing large vocabulary conversational speech recognition (LVCSR) systems,” when many such systems were multilingual and “the need also exists for a trained acoustic model to be easily adapted in structure and size to unseen domains using only small amount of adaptation data,” would be motivated to seek out a system that builds a multilingual or different language speech set when the new domain is a new language, as disclosed by Sabourin. A person having ordinary skill in the art would have a reasonable expectation of success in implementing the teaching of Waibel to generate a multilingual or new language speech recognizer as provided by Sabourin because the combination at least involves the predictable use of prior art elements according to their established functions. Further, recognizers that re-estimated by adding nodes to the second speech recognizer were known. Sabourin and Schultz similarly disclose adding nodes. For the reasons above, a POSITA would similarly be motivated to combine either Sabourin or Schultz with Waibel to create a multilingual speech recognizer where re-estimation includes the addition of nodes to account for new phones or contexts not present in the first speech recognizer. • *See* Exhibit A-2, Claim 1.a.1. • *See* Exhibit A-4, Claim 1.a.1.

APPENDIX A-4
Invalidity Claim Chart for U.S. Pat. No. 6,999,925 ('925 patent)
Schultz, et al., *Japanese LVCSR on the Spontaneous Scheduling Task with JANUS-3*, Eurospeech, Rhodes 1997 ("Schultz")[1]

On October 16, 2020, Nuance narrowed the asserted '925 patent claims to 1, 14 and 27. Nuance's Disclosure Limiting Asserted Claims of Phase 1 Patents. Claims 1, 14 and 27 of the '925 patent are anticipated and/or rendered obvious by Schultz alone or in combination with at least the following references. The chart below discloses how prior art references identified by Omilia disclose, either expressly or inherently, and/or render obvious each of the asserted claims:

(1) U.S. Pat. No. 6,912,499, Sabourin et al., filed August 31, 1999 ("Sabourin")[2]

(2) U.S. Pat. No. 6,324,510, Waibel et al., filed November 6, 1998 ("Waibel")[3]

Schultz incorporated the following prior-art references:

- M. Finke, et al., *Wide Context Acoustic Modeling in Read vs Spontaneous Speech*, Proc. Of ICASSP, Munich 1997 ("Finke")

This invalidity claim chart is based on Omilia's present understanding of the asserted claims, the Court's August 6, 2020 Claim Construction Order (ECF. No. 157), Nuance's initial and all subsequent supplemental Infringement Contentions, its July 7, 2020 Response to Omilia's Supplemental Non-Infringement and Invalidity Responses, Nuance's Response to Omilia's Interrogatory No. 9, and Omilia's investigation to date. Omilia is not adopting Nuance's apparent constructions, nor is Omilia admitting to the accuracy of any particular construction. Omilia reserves all rights to amend this invalidity claim chart if Nuance further amends its Infringement Contentions, or in response to any statements made by Nuance at any phase of this litigation. In addition, Discovery is still ongoing, and Omilia reserves its right to supplement its contentions with additional evidence as discovery continues.

The citations provided are exemplary and do not necessarily include each and every disclosure of the limitation in the references. Omilia has cited the most relevant portions of the identified prior art. Other portions of the identified prior art may additionally disclose, either expressly or inherently, and/or render obvious one or more limitations of the asserted claims. To provide context or aid in understanding the prior art and the state of the art at the time of the alleged invention, Omilia reserves the right to rely on: (1) uncited portions of the

---

[1] Schultz was presented at Eurospeech 97 from September 22-25, 1997 and constitutes prior art at least under 35 U.S.C. § 102(a) & 102(b).

[2] Sabourin was filed on August 31, 1999 and constitutes prior art at least under 35 § 102(a) & 102(e).

[3] Waibel was filed on November 6, 1998 and is prior art at least under 35 U.S.C. § 102(a) & 102(e).

identified prior art; (2) other prior art not identified herein; (3) references that show the state of the art (irrespective of whether such references themselves qualify as prior art to the asserted patents); (4) factual testimony from the inventors or authors of the prior art references, or purveyors of prior art devices; and/or (5) expert testimony.

Citations to a particular drawing or figure in the accompanying charts should be understood to include the description of the drawing or figure, as well as any text associated with the drawing or figure. Citations to particular text concerning a drawing or figure, should also be understood to encompasses that drawing or figure as well. The identified prior art expressly and/or inherently discloses the features of the asserted claims under all proposed constructions of the asserted claims. To establish the inherency of certain features of the prior art to invalidate the asserted claims, Omilia may rely on cited or uncited portions of the prior art, other documents, factual testimony, and expert testimony. To the extent that the Court finds that this reference does not explicitly teach certain limitations in the asserted claims, such limitations would have been inherent and/or obvious and Omilia reserves the right to supplement or amend this claim chart to address such a finding.

’925 Patent		
Claim 1		
1.pre.a	“A computerized method of automatically generating from a first speech recognizer a second speech recognizer”	Schultz discloses a computerized method of automatically generating from a first speech recognizer a second speech recognizer. Schultz describes creating a Japanese speech-to-speech translator by using a previously constructed German speech recognizer. “This paper presents our findings during development of the recognition engine for the Japanese art of the VERBMOBIL speech-to-speech translation project. We describe an efficient method to bootstrap a large vocabulary speech recognizer for spontaneously spoken Japanese from a German recognizer and show that the amount of effort in developing the system could be reduced by using this rapid cross language bootstrapping technique. The Japanese recognizer is integrated into the VERBMOBIL system and shows very promising results achieving 9.3% word error rate.” Schultz at Abstract. “To bootstrap the Japanese system we took a German context-independent 3-state HMM recognizer. Each state of the HMM is modeled by one codebook. Each codebook contains 16 mixture Gaussian distribution of a 32 dimensional feature space. 16 Mel-scale coefficients, power and their first and second derivatives are calculated from the 16

’925 Patent		
		kHz sampled input speech. Mean subtraction is applied. The amount of features is reduced to 32 coefficients by computing a Linear Discrimination Analysis (LDA). We created a copy of the German recognizer containing the codebooks and distributions of the German counterparts for the Japanese models as described in table 2. Additional copies for codebooks and distributions from similar German phonemes are added for the 4 phonemes not found in the German phoneme set. As can be seen in table 3 this recognizer has a reasonable performance, though it has never seen any Japanese data. In the next step this bootstrapped version was trained 3 iterations. A new LDA had been calculated, and new codebooks were clustered. Another 3 training iteration made out the context-independent Japanese system, which leads to 20.9% word error rate.” Schultz at Section 3.3. To the extent, Schultz does not explicitly describe the method as automatic, a POSITA would have understood that the steps of automating are described as conventional and common place. The ’925 patent concedes this in its description and use of computer readable storage medium. ’925 patent at 3:29-33, 9:48-53. Schultz likewise references the end implementation in a computer system, VERMOBIL. Moreover, the use of computers to implement this method is obvious. Not simply by Schultz, but also in light of Sabourin and Waibel. It would be obvious to a person having ordinary skill in the art to combine Schultz with Sabourin or Waibel to disclose automatically generate the second decision network. Schultz, Sabourin, and Waibel are in the same field of art. A person having ordinary skill in the art would be motivated to combine Schultz with Sabourin or Waibel. A person having ordinary skill in the art considering Schultz’s disclosure that “the [human performed] segmentation approach . . . results in relatively small vocabulary growth rates as compared to the non-segmented data” would be motivated to seek out a system that adapts the first decision network automatically to achieve improved results, as disclosed by Waibel, to attempt to achieve even further improved system. A person having ordinary skill in the art would have a reasonable expectation of success in implementing the

’925 Patent		
		teaching of Schultz to automatically generate the second decision network as provided by Waibel because the combination involves the predictable use of prior art elements according to their established functions. “In another embodiment, the labels can be assigned using a speech recognizer unit by Selecting the top scoring recognition candidate as the label for the utterance processed by the recognizer.” Sabourin at 6:12-15. “This invention relates to speech model sets and to a method and apparatus for training speech model sets for use in speech recognition systems operating in multilingual environments as may be used in a telephone directory assistance system, voice activated dialing (VAD) system, personal voice dialing systems and other speech recognition enabled services.” “The present invention is directed to a method of organizing an acoustic model for speech recognition comprised of the steps of calculating a measure of acoustic dissimilarity of subphonetic units. Recursively clustering the subphonetic units based on the calculated measure automatically generates a hierarchically arranged model. An apparatus for performing the method is also disclosed.” Waibel, at 2:52–58. “Starting from an initial set of decision tree clustered, context-dependent, subphonetic units, the present invention uses an aglommerative clustering algorithm across monophones to automatically design a tree-structured decomposition of posterior probabilities which is instantiated with thousands of small neural network estimators at each of the nodes of the tree.” Waibel, at 2:59–65.

’925 Patent		
1.pre.b	“said first speech recognizer comprising a first acoustic model with a first decisions network and corresponding first phonetic contexts”	Schultz discloses a first speech recognizer with a first acoustic model, first decision network and corresponding phonetic contexts. Schultz describes a German language recognizer using a German dictionary, phoneme codebook, and polyphone decision tree using HMM. *See, e.g.*, “To bootstrap the Japanese system we took a German context-independent 3-state HMM recognizer. Each state of the HMM is modeled by one codebook. Each codebook contains 16 mixture Gaussian distribution of a 32 dimensional feature space. 16 Mel-scale coefficients, power and their first and second derivatives are calculated from the 16 kHz sampled input speech. Mean subtraction is applied. The amount of features is reduced to 32 coefficients by computing a Linear Discrimination Analysis (LDA). We created a copy of the German recognizer containing the codebooks and distributions of the German counterparts for the Japanese models as described in table 2. Additional copies for codebooks and distributions from similar German phonemes are added for the 4 phonemes not found in the German phoneme set. As can be seen in table 3 this recognizer has a reasonable performance, though it has never seen any Japanese data. In the next step this bootstrapped version was trained 3 iterations. A new LDA had been calculated, and new codebooks were clustered. Another 3 training iteration made out the context-independent Japanese system, which leads to 20.9% word error rate.” Schultz at Section 3.3. “After training and testing the context-independent system a context-dependent system was developed based on the JANUS-3 toolkit. Analyzing the scheduling database shows that only 54000 different sept-phones (a context of 3 phonemes to the right and to the left) could be found. The Japanese phonetic seems to be more restricted compared to German speech in the equivalent appointment scheduling task (200.000 quint-phones) and for English in the switch-board task (500.000 triphones).

’925 Patent		
		We modeled 54000 sept-phones in the context-dependent Japanese system and clustered these sept-phones to 600 decision-tree-clustered polyphone models as described in [8]. The final context-dependent system has about 2000 distributions over the 600 polyphone models. The phonetic questions needed for the clustering procedure could be derived from the German phonetic questions. The resulting contextdependent Japanese speech recognition system achieves 13.0% word error rate.” Schultz at Section 3.4. “Using the full data-set for the estimation of the HMM-parameters we changed the system structure to a fully continuous approach with an increased number of codebooks. The polyphonic tree of all occurring sept-phones (containing cross-word models with up to one phoneme lookahead to adjacent words) has been clustered to 2000 codebooks =, each of which has been modeled as a mixture of 32 Gaussians with diagonal covariance. In order to increase recognition speed the dimensionality of the feature set was reduced to the first 24 LDA parameters of the feature set described in section 3.3. Label boosting (using supervised MLLR-adaption) was used to improve the quality of the database labeling.” Schultz at Section 4.1. To the extent, Schultz does not explicitly disclose a first acoustic model with first decision network and first phonetic context, a POSITA would have recognized that HMMs naturally are organized in decision networks and a typical type of network is a decision tree. Further each state in an HMM system is the resulting state of the decision network, meaning that the state is a state in context to the rest of the decision network. This context is the phonetic context. A POSITA would have recognized these properties of an HMM system as described in Waibel. “The disclosed method allows effective adaptation of the structure and size of a trained acoustic model to an unseen domain using only a small amount of adaptation data. The

’925 Patent		
		present invention benefits from the multi-level, hierarchical representation of the context-dependent acoustic model. In contrast to approaches based on acoustic adaptation only, the present invention uses an estimate of the a-priori distribution of modeled HMM states on the new domain to dynamically downsize or prune the tree-structured acoustic model. In that manner, the present invention accounts for differences in vocabulary size and adjusts to the specificity of phonetic context observed in the new domain.” Waibel, at 3:11–22.
1.pre.c	“and said second speech recognizer being adapted to a specific domain, said method comprising:”	Schultz discloses that the second speech recognizer is adapted to a specific domain. Schultz discloses creating a Japanese speech recognizer from the German speech recognizer. *See, e.g.*, “This paper presents our findings during development of the recognition engine for the Japanese art of the VERBMOBIL speech-to-speech translation project. We describe an efficient method to bootstrap a large vocabulary speech recognizer for spontaneously spoken Japanese from a German recognizer and show that the amount of effort in developing the system could be reduced by using this rapid cross language bootstrapping technique. The Japanese recognizer is integrated into the VERBMOBIL system and shows very promising results achieving 9.3% word error rate.” Schultz at Abstract. “To bootstrap the Japanese system we took a German context-independent 3-state HMM recognizer. Each state of the HMM is modeled by one codebook. Each codebook contains 16 mixture Gaussian distribution of a 32 dimensional feature space. 16 Mel-scale coefficients, power and their first and second derivatives are calculated from the 16 kHz sampled input speech. Mean subtraction is applied. The amount of features is reduced to 32 coefficients by computing a Linear Discrimination Analysis (LDA). We created a copy of the German recognizer containing the codebooks and distributions of the German counterparts for the Japanese models as described in table 2. Additional

’925 Patent		
		copies for codebooks and distributions from similar German phonemes are added for the 4 phonemes not found in the German phoneme set. As can be seen in table 3 this recognizer has a reasonable performance, though it has never seen any Japanese data. In the next step this bootstrapped version was trained 3 iterations. A new LDA had been calculated, and new codebooks were clustered. Another 3 training iteration made out the context-independent Japanese system, which leads to 20.9% word error rate.” Schultz at Section 3.3.
1.a.1	“based on said first acoustic model, generating a second acoustic model with a second decision network and corresponding second phonetic contexts for said second speech recognizer by re-estimating said first decision network and said corresponding first phonetic contexts based on domain-specific training data,”	Schultz discloses that based on said first acoustic model, generating a second acoustic model with a second decision network and corresponding second phonetic contexts for said second speech recognizer by re-estimating said first decision network and said corresponding first phonetic contexts based on domain-specific training data. Schultz discloses training the new system with the Japanese data such that the decision tree is re-clustered to be context dependent and select those codebooks of the German recognizer that are used in the Japanese language while adding new codebooks for Japanese phones that are not in the German language. *See, e.g.*, “This paper presents our findings during development of the recognition engine for the Japanese art of the VERBMOBIL speech-to-speech translation project. We describe an efficient method to bootstrap a large vocabulary speech recognizer for spontaneously spoken Japanese from a German recognizer and show that the amount of effort in developing the system could be reduced by using this rapid cross language bootstrapping technique. The Japanese recognizer is integrated into the VERBMOBIL system and shows very promising results achieving 9.3% word error rate.” Schultz at Abstract. “To bootstrap the Japanese system we took a German context-independent 3-state HMM recognizer. Each state of the HMM is modeled by one codebook. Each codebook contains 16 mixture Gaussian distribution of a 32 dimensional feature space. 16 Mel-scale coefficients, power and their first and second derivatives are calculated from the 16

’925 Patent		
		kHz sampled input speech. Mean subtraction is applied. The amount of features is reduced to 32 coefficients by computing a Linear Discrimination Analysis (LDA). We created a copy of the German recognizer containing the codebooks and distributions of the German counterparts for the Japanese models as described in table 2. Additional copies for codebooks and distributions from similar German phonemes are added for the 4 phonemes not found in the German phoneme set. As can be seen in table 3 this recognizer has a reasonable performance, though it has never seen any Japanese data. In the next step this bootstrapped version was trained 3 iterations. A new LDA had been calculated, and new codebooks were clustered. Another 3 training iteration made out the context-independent Japanese system, which leads to 20.9% word error rate.” Schultz at Section 3.3. “After training and testing the context-independent system a context-dependent system was developed based on the JANUS-3 toolkit. Analyzing the scheduling database shows that only 54000 different sept-phones (a context of 3 phonemes to the right and to the left) could be found. The Japanese phonetic seems to be more restricted compared to German speech in the equivalent appointment scheduling task (200.000 quint-phones) and for English in the switch-board task (500.000 triphones). We modeled 54000 sept-phones in the context-dependent Japanese system and clustered these sept-phones to 600 decision-tree-clustered polyphone mo- dels as described in [8]. The final context-dependent system has about 2000 distributions over the 600 polyphone models. The phonetic questions needed for the clustering procedure could be derived from the German phonetic questions. The resulting context-dependent Japanese speech recognition system achieves 13.0% word error rate.” Schultz at Section 3.4.

’925 Patent		
		Further, recognizers that re-estimated the first recognizer by deleting or merging nodes were known. For example, Sabourin describes a similar process with the added process of removing nodes below a certain threshold and merging other nodes where desired. Moreover, given this disclosure, the re-estimation of the first speech recognizer was obvious in light of Schultz alone, or in combination with Sabourin. The motivation to combine these references would at least include: • Combining prior art elements according to known methods to yield predictable results (deleting unused phones were known using similar techniques for predictable results); • Simple substitution of one known element for another to obtain predictable results; • Use of known technique to improve similar devices (methods, or products) in the same way (same technique of adapting the decision network when expanding the domain of the recognizer to include new phones or deleting unused phones was known)); • Applying a known technique to a known device (method, or product) ready for improvement to yield predictable results (deleting nodes in the decision network during adaptation were known using similar known techniques for predictable results); • “Obvious to try”—choosing from a finite number of identified, predictable solutions, with a reasonable expectation of success (obvious to try to remove nodes for unused phones for better recognition accuracy); • Market forces and benefits associated with the known benefits of deleting nodes were predictable to POSA • Teaching of prior art would have lead a POSA to combine the references to arrive at an adaptation of a recognizer that includes deleting nodes.

’925 Patent		
		For example, Sabourin explicitly re-estimates by adding nodes, pruning nodes and merging nodes in the decision network. It would be obvious to a person having ordinary skill in the art to combine Schultz with Sabourin to disclose adding nodes, pruning nodes and merging nodes in the decision network. Both Schultz and Sabourin are in the same field of art. A person having ordinary skill in the art would be motivated to combine Schultz with Sabourin. A person having ordinary skill in the art considering Schultz’s disclosure that “the word error rate of 9.3% [of the resulting system] achieves very promising results” would be motivated to seek out a system that adapts the first decision network by merging and deleting nodes, as disclosed by Sabourin, to attempt to achieve even further improved system. A person having ordinary skill in the art would have a reasonable expectation of success in implementing the teaching of Schultz to re-estimate the first decision network through addition of nods, merging of nodes and deleting of nodes as provided by Sabourin because the combination involves the predictable use of prior art elements according to their established functions. *See* Sabourin at Abstract, 2:57-3:6, 3:52-4:7, 4:16-45, 5:66-6:44, 6:45-7:8, 8:49-10:23.
1.a.2	“wherein said first decision network and said second decision network utilize a phonetic decision free [sic] to perform speech recognition operations”	Schultz discloses that the first and second decision networks utilized phonetic decision trees to perform speech operations. In Schultz, both the Japanese and the German recognizers use polyphone decision trees as the decision network. *See* claim 1.pre.b. *See also:* “Using the full data-set for the estimation of the HMM-parameters we changed the system structure to a fully continuous approach with an increased number of codebooks. The polyphonic tree of all occurring sept-phones (containing cross-word models with up to one phoneme lookahead to adjacent words) has been clustered to 2000 codebooks, each of which has been modeled as a mixture of 32 Gaussians with diagonal covariance. In order to increase recognition speed the dimensionality of the feature set was reduced to the first 24 LDA parameters of the feature set described in section 3.3. Label boosting (using supervised MLLR-adaption) was used to improve the quality of the database labeling.”

’925 Patent		
		Schultz at Section 4.1. “To bootstrap the Japanese system we took a German context-independent 3-state HMM recognizer. Each state of the HMM is modeled by one codebook. Each codebook contains 16 mixture Gaussian distribution of a 32 dimensional feature space. 16 Mel-scale coefficients, power and their first and second derivatives are calculated from the 16 kHz sampled input speech. Mean subtraction is applied. The amount of features is reduced to 32 coefficients by computing a Linear Discrimination Analysis (LDA). We created a copy of the German recognizer containing the codebooks and distributions of the German counterparts for the Japanese models as described in table 2. Additional copies for codebooks and distributions from similar German phonemes are added for the 4 phonemes not found in the German phoneme set. As can be seen in table 3 this recognizer has a reasonable performance, though it has never seen any Japanese data. In the next step this bootstrapped version was trained 3 iterations. A new LDA had been calculated, and new codebooks were clustered. Another 3 training iteration made out the context-independent Japanese system, which leads to 20.9% word error rate.” Schultz at Section 3.3. “After training and testing the context-independent system a context-dependent system was developed based on the JANUS-3 toolkit. Analyzing the scheduling database shows that only 54000 different sept-phones (a context of 3 phonemes to the right and to the left) could be found. The Japanese phonetic seems to be more restricted compared to German speech in the equivalent appointment scheduling task (200.000 quint-phones) and for English in the switch-board task (500.000 triphones). We modeled 54000 sept-phones in the context-dependent Japanese system and clustered these sept-phones to 600 decision-tree-clustered polyphone mo-

’925 Patent		
		dels as described in [8]. The final context-dependent system has about 2000 distributions over the 600 polyphone models. The phonetic questions needed for the clustering procedure could be derived from the German phonetic questions. The resulting context-dependent Japanese speech recognition system achieves 13.0% word error rate.” Schultz at Section 3.4. Schultz specifically contemplates that the first and second decision networks will be polyphonic decision trees as those in Finke. Schultz references Finke in saying “We modeled 54000 sept-phones in the context-dependent Japanese system and clustered these sept-phones to 600 decision-tree-clustered polyphone models as described in [8],” in that reference 8 is Finke. Finke describes using a polyphonic clustering algorithm to collect polyphones and organize them into a decision tree. “Due to the extremely large number of models we have to handle within the clustering procedures, we had to come up with efficient data structures to organize the polyphones and their associated distributions. One efficient way to represent a set of polyphones are so called polyphonic trees: The root of the tree is the centric/mid phone. For each observed immediate context +-1 there is a child to the root node with the names of the left and the right phone, the count of how often the respective “triphone” was observed, and a pointer to the acoustic model (distribution). Each “triphone” child has a set of children one for each “quintphone” context found around the triphone parent in the training data.” Finke at Section 3.1, *see also* Fig. 4.

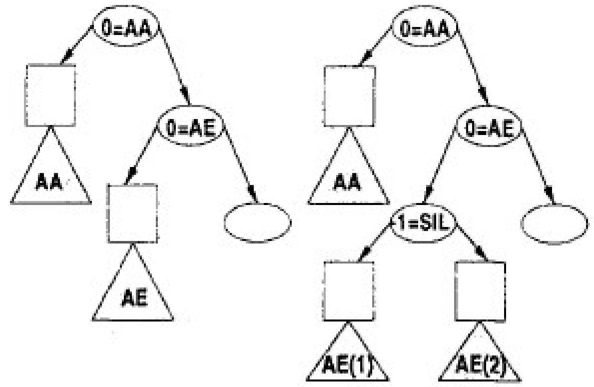
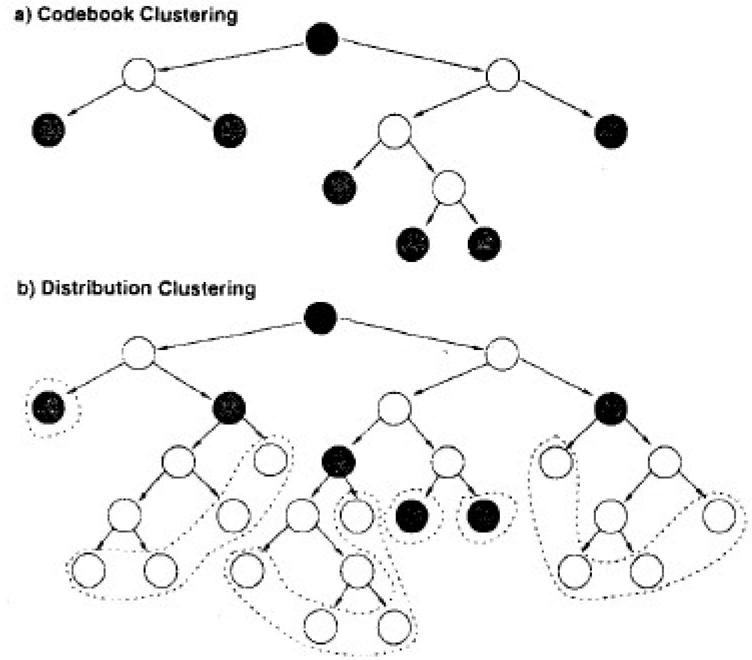
<table>
<tr><th colspan="3"><u>’925 Patent</u></th></tr>
<tr><td></td><td></td><td>

Figure 4. Splitting a decision tree node and its associated polyphonic tree based on a phonetic question
“In our standard training scheme we first grow a decision tree until it reaches the number of desired leaf nodes (typically a few thousand, depending on the size of the available training data; grey nodes in Figure 1). We constraint splits to be only valid as long as both child nodes created still have sufficient training data to train the underlying codebook. Then, a fully continuous Gaussian mixture model is trained for every leaf node. In a second clustering phase, we continue growing the decision tree and eventually train a separate distribution of mixture weights for each of the resulting leafs. This is a new way of optimizing the degree of mixture tying in a large vocabulary hidden Markov model based speech recognition system.”
Finke at Section 3.4, see also Tab. 1. </td></tr>
</table>

’925 Patent		
		 Table 1. Two stage clustering of acoustic models (the distributions in the same dashed area are defined on the same codebook) “Hidden Markov models with continuous densities provide a detailed stochastic representation of the acoustic space at the expense of increased computational complexity and lack of robustness. This two level clustering approach addresses the problem of the lack of robustness by having a set of distributions share the same codebook. In particular, the algorithm proposed hr4ps to automatically determine the number of sets of HMM states which share the same codebook and based on that subsets of HMM states which share the same distribution.” Finke at Section 3.4.

’925 Patent		
		Finke is incorporated by the extensive reference and use in Schultz. However, to the extent a POSITA would not find the incorporation explicit, it would be obvious to a POSITA to combine Schultz with Finke to disclose the use of a phonetic decision tree to recognize speech. Both Schultz and Finke are in the same field of art. A person having ordinary skill in the art would be motivated to combine Schultz with Rinke. A person having ordinary skill in the art considering Schultz’s disclosure that “the context-dependent Japanese system and clustered these sept-phones to 600 decision-tree-clustered polyphone models” would be motivated to seek out a reference that explicitly describes the use of a decision tree for recognizing speech, as disclosed by Finke. A person having ordinary skill in the art would have a reasonable expectation of success in implementing the teaching of Schultz with the use of decision trees as provided by Finke because the combination involves the predictable use of prior art elements according to their established functions.
1.b.1	“wherein the number of nodes in the second decision network is not fixed by the number of nodes in the first decision network,”	Schultz discloses that the number of nodes in the second decision network is not fixed by the number of nodes in the first decision network. Schultz teaches a method that copies German codebooks corresponding to Japanese phones and then adds clusters and code books for Japanese phonemes that are not present in the German phoneme sets and then also re-clustering the decision tree based on training data. “We created a copy of the German recognizer containing the codebooks and distributions of the German counterparts for the Japanese models as described in table 2. Additional copies for codebooks and distributions from similar German phonemes are added for the 4 phonemes not found in the German phoneme set. As can be seen in table 3 this recognizer has a reasonable performance, though it has never seen any Japanese data. In the next step this bootstrapped version was trained 3 iterations. A new LDA had been calculated, and new codebooks were clustered. Another 3 training iteration made out the context-independent Japanese system, which leads to 20.9% word error rate.” Schultz at Section 3.3; *see also* Schultz Table 2.

’925 Patent		
1.b.2	“and wherein said re-estimating comprises partitioning said training data using said first decision network of said first speech recognizer.”	Schultz discloses that the re-estimation comprises partitioning said training data using said first decision network of said first speech recognizer. In Schultz, the data from the Japanese or German language is mapped to the phones of the first language. “Running an alignment of a German phoneme recognizer on Japanese input speech gave us the impression, that the Japanese phonetic is very similar to German. Therefore we decided to bootstrap the Japanese phoneme set with German models developed for the German VERBMOBIL recognizer system. Only 4 of the 31 phonemes required for acoustic modeling have no German counterparts. These were bootstrapped as follows:/4/ from /u/, /4:/ from /u:/, /&0/ from /s/, and /dZ/ from /tS/. To cope with the effects arising in the spontaneously spoken data, like i.e. stuttering, false starts, or mumbling, special noise models [7] were included. All together 44 different phonemes are used to model Japanese speech: 31 speech models, 11 noise models, 1 silence and 1 glottal stop (ref. table 2). After training a Japanese system we determined the similarity between the Japanese and German phoneme set by performing the following experiment: we trained a SCHMM for the combined phoneme set with both German and Japanese input and ran a clustering procedure. This leads to the result that the most similar phonemes in this order are the consonant /z/ and /b/, the affricate /ts/ and the semi vowel /j/. Japanese short vowels are similar to the long version of their German counterparts.” Schultz at Section 3.1, s*ee* Table 2. “To bootstrap the Japanese system we took a German context-independent 3-state HMM recognizer. Each state of the HMM is modeled by one codebook. Each codebook contains 16 mixture Gaussian distribution of a 32 dimensional feature space. 16 Mel-scale coefficients, power and their first and second derivatives are calculated from the 16 kHz sampled input speech. Mean subtraction is applied. The amount of features is reduced to 32 coefficients by computing a Linear Discrimination Analysis (LDA). We created a copy of the German recognizer containing the codebooks and distributions of the German counterparts for the Japanese models as described in table 2. Additional

’925 Patent		
		copies for codebooks and distributions from similar German phonemes are added for the 4 phonemes not found in the German phoneme set. As can be seen in table 3 this recognizer has a reasonable performance, though it has never seen any Japanese data. In the next step this bootstrapped version was trained 3 iterations. A new LDA had been calculated, and new codebooks were clustered. Another 3 training iteration made out the context-independent Japanese system, which leads to 20.9% word error rate.” Schultz at Section 3.3.
Claim 14		
14.pre	A machine-readable storage, having stored thereon a computer program having a plurality of code sections executable by a machine for causing the machine to automatically generate from a first speech recognizer a second speech recognizer, said first speech recognizer comprising a first acoustic model with a first decision network and corresponding first phonetic contexts, and said second speech recognizer being adapted to a specific domain, said machine-readable storage causing the machine to perform the steps of:	*See* claim 1, including 1.pre.a-c. Schultz is implemented with a computer, which has these elements. In any case, it is obvious in view of Sabourin and Waibel which describe computer readable storage medium and apparatus with software capable of performing the method of claim 1. For the reasons described in claim 1 and 27, a POSITA would be motivated to combine Schultz with either one of those references.

’925 Patent		
14.a	based on said first acoustic model, generating a second acoustic model with a second decision network and corresponding second phonetic contexts for said second speech recognizer by re-estimating said first decision network and said corresponding first phonetic contexts based on domain-specific training data, wherein said first decision network and said second decision network utilize a phonetic decision tree to perform speech recognition operations,	*See* claim 1, including 1.a.1-2, and 14.pre.
14.b	wherein the number of nodes in the second decision network is not fixed by the number of nodes in the first decision network, and wherein said re-estimating comprises partitioning said training data using said first decision network of said first speech recognizer.	*See* claim 1, including 1.b.1-2, and 14.pre, 14.a.
Claim 27		
27.pre	“A computerized method of generating a second speech	*See* claim 1, including 1.pre.a-c.

’925 Patent		
	recognizer comprising the steps of:	
27.a	identifying a first speech recognizer of a first domain comprising a first acoustic model with a first decision network and corresponding first phonetic contexts;”	*See* claim 1, including 1.pre.b and 27.pre.
27.b	“receiving domain-specific training data of a second domain; and”	Schultz discloses receiving domain-specific training data of a second domain. See claim 1 including 1.pre.c and 1.a.1. In Schultz, Japanese training data was collected from different native speakers and is used in creating the Japanese recognizer. *See, e.g.*, “The Japanese part of the VERBMOBIL database consists of 800 dialogs spoken by 324 different native speakers. On average the dialogs cover 14 utterances each of a length of about 30 words. All dialogs are collected by ATR Interpreting Telecommunication Laboratories and the University of Electro-Communications in Tokyo (Japan). Further information about the corpus and collection procedures are given in [4].” Schultz at Section 2. “We created a copy of the German recognizer containing the codebooks and distributions of the German counterparts for the Japanese models as described in table 2. Additional copies for codebooks and distributions from similar German phonemes are added for the 4 phonemes not found in the German phoneme set. As can be seen in table 3 this recognizer has a reasonable performance, though it has never seen any Japanese data. In the next step this bootstrapped version was trained 3 iterations. A new LDA had been calculated, and new codebooks were clustered. Another 3 training iteration made out the context-independent Japanese system, which leads to 20.9% word error rate.” Schultz at Section 3.3.

’925 Patent		
27.c	“based on the first speech recognizer and the domain-specific training data, generating a second acoustic model of said first domain and said second domain comprising a second acoustic model with a second decision network and corresponding second phonetic contexts, wherein the first domain comprises at least a first language, wherein the second domain comprises at least a second language, and wherein the second speech recognizer is a multi-lingual speech recognizer.”	Schultz discloses “based on the first speech recognizer and the domain-specific training data, generating a second acoustic model of said first domain and said second domain comprising a second acoustic model with a second decision network and corresponding second phonetic contexts.” In Schultz, the first (German language) recognizer is adapted into the second (Japanese language) recognizer. *See* claim 1 including 1.pre.c and 1.a.1. Schultz does not explicitly contemplate that the final speech recognizer will be able to understand both the Japanese and German languages, Schultz contemplates that its process permits a fast adaptation of a known speech recognizer to a new language with reasonable results and reduced amount of effort. *See* Schultz at Abstract. However, the Japanese system would implicitly be able to understand both the Japanese language and the German language so long as both dictionaries are retained. Further, a POSITA would understand that this same adaptation could be performed with a pre-existing multi-lingual recognizer or used to create a multilingual recognizer. This was a common application and need for modified recognizers. For example, multilingual speech recognizers were well known at the time of the ’925 patent. *See, e.g.*, Schultz et al., “*Polyphone Decision Tree Specialization for Language Adaptation*”, ICASSP-2000, Istanbul, Turkey, Jun. 2000 (Section 2.2. describing generating and use of multilingual speech recognizers). A POSITA would have known that multilingual speech recognizers were well known and desirable. *See* Exhibit A-1, (Sabourin Chart, claim 27.c). Moreover, given this disclosure, the creation of a multilingual second recognizer was obvious in light of Schultz alone, or in combination with Sabourin. The motivation to combine these references would at least include: • Combining prior art elements according to known methods to yield predictable results (multilingual recognizers were known using similar techniques for predictable results);

’925 Patent		
		• Simple substitution of one known element for another to obtain predictable results (substituting the domain from just one language to two); • Use of known technique to improve similar devices (methods, or products) in the same way (same technique of expanding domain of the recognizer to include multiple languages was known)); • Applying a known technique to a known device (method, or product) ready for improvement to yield predictable results (multilingual recognizers were known using similar known techniques for predictable results); • “Obvious to try”—choosing from a finite number of identified, predictable solutions, with a reasonable expectation of success (obvious to try to recognize more than one language); • Market forces and benefits associated with the known benefits of multilingual recognizers were predictable to POSA • Teaching of prior art would have lead a POSA to combine the references to arrive at a multilingual recognizer. For example, Sabourin explicitly discloses a multilingual system. Sabourin discloses a multilingual system and where there are two different language domains as recited. It would be obvious to a person having ordinary skill in the art to combine Schultz with Sabourin. Both Schultz and Sabourin are in the same field of art. A person having ordinary skill in the art would be motivated to combine Schultz with Sabourin. A person having ordinary skill in the art considering Schultz’s disclosure that “the Japanese recognizer is integrated into the VERBMOBIL system” and “the VERBMOBILE project is to build speech-to-speech translation system from both German and Japanese spontaneously spoken input speech to English, German and Japanese output in an appointment scenario” would be motivated to seek out a system that builds a multilingual speech set to better help with this translation, as disclosed by Sabourin. A

’925 Patent		
		person having ordinary skill in the art would have a reasonable expectation of success in implementing the teaching of Schultz to generate a multilingual speech recognizer as provided by Sabourin because the combination involves the predictable use of prior art elements according to their established functions. “The invention relates to a method and apparatus for training a multilingual speech model Set. The multilingual Speech model Set generated is Suitable for use by a speech recognition System for recognizing spoken utterances for at least two different languages. The invention allows using a single Speech recognition unit with a single Speech model Set to perform Speech recognition on utterances from two or more languages. The method and apparatus make use of a group of a group of acoustic Sub-word units comprised of a first Subgroup of acoustic Sub-word units associated to a first language and a Second Subgroup of acoustic Sub-word units associated to a Second language where the first Subgroup and the Second Subgroup share at least one common acoustic Sub-word unit. The method and apparatus also make use of a plurality of letter to acoustic Sub-word unit rules Sets, each letter to acoustic Sub-word unit rules Set being associated to a different language. A Set of untrained speech models is trained on the basis of a training Set comprising speech tokens and their associated labels in combination with the group of acoustic Sub-word units and the plurality of letter to acoustic Sub-word unit rules Sets. The invention also provides a computer readable Storage medium comprising a program element for implementing the method for training a multilingual Speech model Set.” Sabourin at Abstract. “A deficiency of the above-described method is that the Speech recognition System requires as an input the language associated to the input utterance, which may not be readily available to the Speech recognition System. Usually, obtaining the language requires prompting the user for the language of use thereby requiring an additionally Step in the Service being provided by the Speech recognition enabled system which may lower the level of satisfaction of the user with the system as a whole. Another deficiency of the above noted method is the costs associated to developing and maintaining a

’925 Patent		
		Speech model Set for each language the Speech recognition System is adapted to recognize. More Specifically, each speech model Set must be trained individually, a task requiring manpower for each individual language thereby increasing significantly the cost of Speech recognition Systems operating in multilingual environments with respect to Systems operating in unilingual environments. In addition, the above-described method requires the Storage of a speech model Set for each language in memory thereby increasing the cost of the Speech recognition System in terms of memory requirements. Finally, the above described method requires testing a speech model Set for each language thereby increasing the testing cost of the Speech recognition System for each language the Speech recognition System is adapted to recognize. Thus, there exists a need in the industry to refine the process of training Speech models So as to obtain an improved multilingual Speech model Set capable of being used by a speech recognition System for recognizing spoken utterances for at least two different languages.” Sabourin at 1:55-2:17. “In accordance with another broad aspect, the invention provides a method for generating a multilingual speech model Set Suitable for use in a multilingual speech recognition System. The method comprises providing a group of acoustic Sub-word units having a first Subgroup of acoustic Sub-word units associated to a first language and a Second Subgroup of acoustic Sub-word units associated to a Second language. The first Subgroup and the Second Subgroup share at least one common acoustic Sub-word unit. The method further comprises providing a training Set comprising a plurality of entries, each entry having a speech token representative of a word and a label being an orthographic representation of the word. The method further comprises providing a Set of untrained speech models and training the Set of untrained speech models by utilizing the training Set, the plurality of letter to acoustic Sub-word unit rules Sets and the group of acoustic Sub-word units to derive the multilingual Speech model Set.” Sabourin at 2:57-3:7.

’925 Patent		
		“The invention provides a method for generating the multilingual speech model set shown in FIG.1. As shown in FIG. 2 of the drawings, the method comprises providing 200 a group of acoustic Sub-word units comprised of a first Subgroup of acoustic Sub-word units associated to a first language and a Second Subgroup of acoustic Sub-word units associated to a Second language. The first Subgroup and the Second Subgroup share at least one common acoustic Sub word unit. In a preferred embodiment, the group of acoustic Sub word unit is obtained by combining sub-word units from individual languages. For each language for which the multilingual Speech model Set is to be representative of, an inventory of sub-word units can be obtained. The acoustic Sub-word units are basic units of Sound in a language. These inventories may be produced by phoneticians or may be provided as off-the-shelf components. In a specific example, the Sub-word units are phonemes. The Skilled person in the art will readily see that other sub-word units may be used here without detracting from the spirit of the invention.” Sabourin at 4:25-45. “The training of the set of untrained speech models further comprises processing 304 the group of transcriptions generated at step 300 on the basis of a speech token of the corresponding entry in the training Set whereby training the Set of untrained speech models to derive the multilingual Speech model Set.” Sabourin at 9:34-39, *see* Figs. 1-7.

APPENDIX A-5
Invalidity Claim Chart for U.S. Pat. No. 6,999,925 ('925 patent)
U.S. Pat. No. 6,789,061 ("Fischer")

On October 16, 2020, Nuance narrowed the asserted '925 patent claims to 1, 14 and 27. Nuance's Disclosure Limiting Asserted Claims of Phase 1 Patents. Claims 1, 14, and 27 of the '925 patent are invalid for obvious type double patenting in view of claims 6 and 15 of the '061 patent alone or in combination with at least the following references. The chart below discloses how prior art references identified by Omilia disclose, either expressly or inherently, and/or render obvious these asserted claims:

(1) U.S. Pat. No. 6,912,499, Sabourin et al., filed August 31, 1999 ("Sabourin")[1]

(2) Schultz, et al., *Japanese LVCSR on the Spontaneous Scheduling Task with JANUS-3*, Eurospeech, Rhodes 1997 ("Schultz")[2]

This invalidity claim chart is based on Omilia's present understanding of the asserted claims, the Court's August 6, 2020 Claim Construction Order (ECF. No. 157), Nuance's initial and all subsequent supplemental Infringement Contentions, its July 7, 2020 Response to Omilia's Supplemental Non-Infringement and Invalidity Responses, Nuance's Response to Omilia's Interrogatory No. 9, and Omilia's investigation to date. Omilia is not adopting Nuance's apparent constructions, nor is Omilia admitting to the accuracy of any particular construction. Omilia reserves all rights to amend this invalidity claim chart if Nuance further amends its Infringement Contentions, or in response to any statements made by Nuance at any phase of this litigation. In addition, Discovery is still ongoing, and Omilia reserves its right to supplement its contentions with additional evidence as discovery continues.

The citations provided are exemplary and do not necessarily include each and every disclosure of the limitation in the references. Omilia has cited the most relevant portions of the identified prior art. Other portions of the identified prior art may additionally disclose, either expressly or inherently, and/or render obvious one or more limitations of the asserted claims. To provide context or aid in understanding the prior art and the state of the art at the time of the alleged invention, Omilia reserves the right to rely on: (1) uncited portions of the identified prior art; (2) other prior art not identified herein; (3) references that show the state of the art (irrespective of whether such references themselves qualify as prior art to the asserted patents); (4) factual testimony from the inventors or authors of the prior art references, or purveyors of prior art devices; and/or (5) expert testimony.

[1] Sabourin was filed on August 31, 1999 and is prior art at least under 35 U.S.C. § 102(a) & 102(e).

[2] Schultz was presented at Eurospeech 97 from September 22-25, 1997 and constitutes prior art at least under 35 U.S.C. § 102 (a) & 102(b).

Citations to a particular drawing or figure in the accompanying charts should be understood to include the description of the drawing or figure, as well as any text associated with the drawing or figure. Citations to particular text concerning a drawing or figure, should also be understood to encompasses that drawing or figure as well. The identified prior art expressly and/or inherently discloses the features of the asserted claims under all proposed constructions of the asserted claims. To establish the inherency of certain features of the prior art to invalidate the asserted claims, Omilia may rely on cited or uncited portions of the prior art, other documents, factual testimony, and expert testimony. To the extent that the Court finds that this reference does not explicitly teach certain limitations in the asserted claims, such limitations would have been inherent and/or obvious and Omilia reserves the right to supplement or amend this claim chart to address such a finding.

’925 Patent		
Claim 1		
1.pre.a	“A computerized method of automatically generating from a first speech recognizer a second speech recognizer”	Fischer discloses automatically generating a second speech recognizer from a first speech recognizer. *See, e.g.,*: “1. **A computer-based method of automatically generating, from a first speech recognizer, a second speech recognizer** wherein the first speech recognizer includes a set of states and a set of probability functions assembling output probabilities for an observation of a speech frame in the states, the method comprising the steps of: generating, from the set of states of the first speech recognizer, a set of states of the second speech recognizer by selecting a subset of states of the first speech recognizer being distinctive of a particular application; and generating, from the set of probability density functions of the first speech recognizer, a set of probability density functions of the second speech recognizer by selecting a subset of probability density functions of the first speech recognizer being distinctive of the particular application, such that the second speech recognizer is at least one tailored to the particular application and requires reduced resources compared to the first speech recognizer. 2. The method of claim 1, further comprising the step of generating acoustic model parameters of the second speech recognizer by re-estimating acoustic model parameters

’925 Patent		
		of the first speech recognizer based on the set of states of the second speech recognizer and the set of probability density functions of the second speech recognizer. 4. The method of claim 2, wherein selecting at least one of the subset of states and the subset of probability density functions of the first speech recognizer exploits phonetical knowledge of the particular application. 5. The method of claim 4, wherein selecting at least one of the subset of states and the subset of probability density functions of the first speech recognizer exploits application specific training data. 6. The method of claim 5, wherein selecting the subset of states comprises associating a multitude of speech frames of the training data with the correct states of the first speech recognizer and selecting those states with a frequency of occurrence above a threshold as the subset of states.” Fischer, claim 6 (emphasis added).
1.pre.b	“said first speech recognizer comprising a first acoustic model with a first decisions network and corresponding first phonetic contexts”	Fischer discloses a first speech recognizer comprising a first acoustic model with a first decision network and corresponding first phonetic contexts. For example, the set of states describes the structure of an acoustic model, where states have a corresponding phonetic context. *See, e.g.,* “1. A computer-based method of automatically generating, from a first speech recognizer, a second speech recognizer **wherein the first speech recognizer includes a set of states and a set of probability functions assembling output probabilities for an observation of a speech frame in the states,** the method comprising the steps of: generating, from the set of states of the first speech recognizer, a set of states of the second speech recognizer by selecting a subset of states of the first speech recognizer being distinctive of a particular application; and

<table>
<tr><th colspan="3"><u>’925 Patent</u></th></tr>
<tr><td></td><td></td><td>generating, from the set of probability density functions of the first speech recognizer, a set of probability density functions of the second speech recognizer by selecting a subset of probability density functions of the first speech recognizer being distinctive of the particular application, such that the second speech recognizer is at least one tailored to the particular application and requires reduced resources compared to the first speech recognizer.

2. The method of claim 1, further comprising the step of generating acoustic model parameters of the second speech recognizer by re-estimating acoustic model parameters of the first speech recognizer based on the set of states of the second speech recognizer and the set of probability density functions of the second speech recognizer.

4. The method of claim 2, wherein selecting at least one of the subset of states and the subset of probability density functions of the first speech recognizer exploits phonetical knowledge of the particular application.

5. The method of claim 4, wherein selecting at least one of the subset of states and the subset of probability density functions of the first speech recognizer exploits application specific training data.

6. The method of claim 5, wherein selecting the subset of states comprises associating a multitude of speech frames of the training data with the correct states of the first speech recognizer and selecting those states with a frequency of occurrence above a threshold as the subset of states.”

Fischer, claim 6 (emphasis added).</td></tr>
<tr><td>1.pre.c</td><td>“and said second speech recognizer being adapted to a specific domain, said method comprising:”</td><td>Fischer discloses that the second speech recognizer is adapted for a specific domain at least because the second speech recognizer is distinctive of a particular application (i.e. domain) See, e.g.,

“1. A computer-based method of automatically generating, from a first speech recognizer, a second speech recognizer wherein the first speech recognizer includes a</td></tr>
</table>

’925 Patent		
		set of states and a set of probability functions assembling output probabilities for an observation of a speech frame in the states, **the method comprising the steps of**: generating, from the set of states of the first speech recognizer, a set of states of the second speech recognizer by selecting a subset of states of the first speech recognizer **being distinctive of a particular application**; and generating, from the set of probability density functions of the first speech recognizer, a set of probability density functions of the second speech recognizer by selecting a subset of probability density functions of the first speech recognizer **being distinctive of the particular application, such that the second speech recognizer is at least one tailored to the particular application** and requires reduced resources compared to the first speech recognizer. 2. The method of claim 1, further comprising the step of generating acoustic model parameters of the second speech recognizer by re-estimating acoustic model parameters of the first speech recognizer based on the set of states of the second speech recognizer and the set of probability density functions of the second speech recognizer. 4. The method of claim 2, wherein selecting at least one of the subset of states and the subset of probability density functions of the first speech recognizer exploits phonetical knowledge of the particular application. 5. The method of claim 4, wherein selecting at least one of the subset of states and the subset of probability density functions of the first speech recognizer exploits application specific training data. 6. The method of claim 5, wherein selecting the subset of states comprises associating a multitude of speech frames of the training data with the correct states of the first speech recognizer and selecting those states with a frequency of occurrence above a threshold as the subset of states.” Fischer, claim 6 (emphasis added).

’925 Patent		
1.a.1	“based on said first acoustic model, generating a second acoustic model with a second decision network and corresponding second phonetic contexts for said second speech recognizer by re-estimating said first decision network and said corresponding first phonetic contexts based on domain-specific training data,”	Fischer discloses based on said first acoustic model, generating a second acoustic model with a second decision network and corresponding second phonetic contexts for said second speech recognizer by re-estimating said first decision network and said corresponding first phonetic contexts based on domain-specific training data at least because Fischer discloses selecting a subset set of states, i.e. deleting nodes form the first decision network. *See, e.g.,:* “1. A computer-based method of automatically generating, from a first speech recognizer, a second speech recognizer wherein the first speech recognizer includes a set of states and a set of probability functions assembling output probabilities for an observation of a speech frame in the states, the method comprising the steps of: **generating, from the set of states of the first speech recognizer, a set of states of the second speech recognizer by selecting a subset of states of the first speech recognizer** being distinctive of a particular application; and **generating, from the set of probability density functions of the first speech recognizer, a set of probability density functions of the second speech recognizer by selecting a subset of probability density functions of the first speech recognizer** being distinctive of the particular application, such that the second speech recognizer is at least one tailored to the particular application and requires reduced resources compared to the first speech recognizer. 2. The method of claim 1, further comprising the step of generating acoustic model parameters of the second speech recognizer by re-estimating acoustic model parameters of the first speech recognizer based on the set of states of the second speech recognizer and the set of probability density functions of the second speech recognizer. 4. The method of claim 2, wherein selecting at least one of the subset of states and the subset of probability density functions of the first speech recognizer exploits phonetical knowledge of the particular application.

’925 Patent		
		5. The method of claim 4, **wherein selecting at least one of the subset of states and the subset of probability density functions of the first speech recognizer exploits application specific training data**. 6. The method of claim 5, wherein selecting the subset of states comprises associating a multitude of speech frames of the training data with the correct states of the first speech recognizer and selecting those states with a frequency of occurrence above a threshold as the subset of states.” Fischer, claim 6 (emphasis added). In addition, speech recognizers that re-estimate by adding nodes was well known at the time of invention. A POSITA would have found the addition of new phonetic contexts and nodes (which is not required by the claims) is an obvious variant to Fischer Claim 6. Moreover, this is obvious in light of Fischer alone or in combination with Sabourin or Schultz. A POSITA would have recognized that Sabourin, Schultz, and Fischer are in similar fields of invention. Moreover, given this disclosure, the re-estimation of the first speech recognizer was obvious in light of Fischer alone, or in combination with Sabourin or Schultz. The motivation to combine these references would at least include: • Combining prior art elements according to known methods to yield predictable results (adding nodes to account for new phones in a domain or deleting unused phones were known using similar techniques for predictable results); • Simple substitution of one known element for another to obtain predictable results; • Use of known technique to improve similar devices (methods, or products) in the same way (same technique of adapting the decision network when expanding the

’925 Patent		
		domain of the recognizer to include new phones or deleting unused phones was known)); • Applying a known technique to a known device (method, or product) ready for improvement to yield predictable results (adding and deleting nodes in the decision network during adaptation were known using similar known techniques for predictable results); • “Obvious to try”—choosing from a finite number of identified, predictable solutions, with a reasonable expectation of success (obvious to add nodes for new sounds or remove nodes for unused phones); • Market forces and benefits associated with the known benefits of adding and deleting nodes were predictable to POSA • Teaching of prior art would have lead a POSA to combine the references to arrive at a adaptation of a recognizer that includes adding or deleting nodes. For example, Sabourin and Schultz explicitly re-estimate by adding nodes, pruning nodes and merging nodes in the decision network. It would be obvious to a POSITA to combine Fischer with Sabourin and Schultz to disclose deleting nodes, adding nodes, pruning nodes and merging nodes in the decision network. Sabourin, Schultz, and Fischer are in the same field of art. A POSITA would be motivated to combine Fischer with Sabourin or Schultz. A POSITA considering Fischer’s disclosure that resources can be re-invested in to the speech recognition system would be motivated to seek out a system that trains the first decision network by adding nodes to account for new sounds, as disclosed by Sabourin and Schultz, which is a way of re-investing nodes. POSITA would have a reasonable expectation of success in implementing the teaching of Fischer to re-estimate the first decision network through addition of nods and deleting of nodes as provided by Sabourin and Schultz because the combination involves the predictable use of prior art elements according to their established functions.

<table>
<tr><th colspan="3"><u>’925 Patent</u></th></tr>
<tr><td></td><td></td><td>See Sabourin at Abstract, 2:57-3:6, 3:52-4:7, 4:16-45, 5:66-6:44, 6:45-7:8, 8:49-10:23; Schultz at Abstract, Section 3.3., Section 3.4.</td></tr>
<tr><td>1.a.2</td><td>“wherein said first decision network and said second decision network utilize a phonetic decision free [sic] to perform speech recognition operations”</td><td>Fischer discloses that the first decision network and second decision network utilize a set of states, which are part of and organized into a phonetic decision tree in speech recognition operations. See, e.g.,

“1. A computer-based method of automatically generating, from a first speech recognizer, a second speech recognizer wherein the first speech recognizer includes a set of states and a set of probability functions assembling output probabilities for an observation of a speech frame in the states, the method comprising the steps of:

generating, from the set of states of the first speech recognizer, a set of states of the second speech recognizer by selecting a subset of states of the first speech recognizer being distinctive of a particular application; and

generating, from the set of probability density functions of the first speech recognizer, a set of probability density functions of the second speech recognizer by selecting a subset of probability density functions of the first speech recognizer being distinctive of the particular application, such that the second speech recognizer is at least one tailored to the particular application and requires reduced resources compared to the first speech recognizer.

2. The method of claim 1, further comprising the step of generating acoustic model parameters of the second speech recognizer by re-estimating acoustic model parameters of the first speech recognizer based on the set of states of the second speech recognizer and the set of probability density functions of the second speech recognizer.

4. The method of claim 2, wherein selecting at least one of the subset of states and the subset of probability density functions of the first speech recognizer exploits phonetical knowledge of the particular application.</td></tr>
</table>

’925 Patent		
		5. The method of claim 4, wherein selecting at least one of the subset of states and the subset of probability density functions of the first speech recognizer exploits application specific training data. 6. The method of claim 5, wherein selecting the subset of states comprises associating a multitude of speech frames of the training data with the correct states of the first speech recognizer and selecting those states with a frequency of occurrence above a threshold as the subset of states.” Fischer, claim 6 (emphasis added).
1.b.1	“wherein the number of nodes in the second decision network is not fixed by the number of nodes in the first decision network,”	Fischer discloses that “the number of nodes in the second decision network is not fixed by the number of nodes in the first decision network.” Fischer describes selecting a subset of states which is necessarily only a subset of the states of the first speech recognizer and therefore nodes are not fixed. *See. e.g.,:* “1. A computer-based method of automatically generating, from a first speech recognizer, a second speech recognizer wherein the first speech recognizer includes a set of states and a set of probability functions assembling output probabilities for an observation of a speech frame in the states, the method comprising the steps of: **generating, from the set of states of the first speech recognizer, a set of states of the second speech recognizer by selecting a subset of states of the first speech recognizer** being distinctive of a particular application; and **generating, from the set of probability density functions of the first speech recognizer, a set of probability density functions of the second speech recognizer by selecting a subset of probability density functions of the first speech recognizer** being distinctive of the particular application, such that the second speech recognizer is at least one tailored to the particular application and requires reduced resources compared to the first speech recognizer.

’925 Patent		
		2. The method of claim 1, further comprising the step of generating acoustic model parameters of the second speech recognizer by re-estimating acoustic model parameters of the first speech recognizer based on the set of states of the second speech recognizer and the set of probability density functions of the second speech recognizer. 4. The method of claim 2, wherein selecting at least one of the subset of states and the subset of probability density functions of the first speech recognizer exploits phonetical knowledge of the particular application. 5. The method of claim 4, wherein selecting at least one of the subset of states and the subset of probability density functions of the first speech recognizer exploits application specific training data. 6. The method of claim 5, wherein selecting the subset of states comprises associating a multitude of speech frames of the training data with the correct states of the first speech recognizer and selecting those states with a frequency of occurrence above a threshold as the subset of states.” Fischer, claim 6 (emphasis added). In addition, Fischer in combination with Sabourin and/or Schultz discloses the number of nodes are not fixed by virtue of the addition or deletion of nodes in the re-estimation process. For the same reasons stated above, a POSITA would be motivated to combine Fischer with Sabourin and Schultz. *See* claim 1.a.1.
1.b.2	“and wherein said re-estimating comprises partitioning said training data using said first decision network of said first speech recognizer.”	Fischer specifically discloses partitioning training data using said first decision network. Specifically, Fischer’s method includes associating the speech frames of the training data set with the correct states of the speech recognizer. *See, e.g.,* “1. A computer-based method of automatically generating, from a first speech recognizer, a second speech recognizer wherein the first speech recognizer includes a set of states and a set of probability functions assembling output probabilities for an observation of a speech frame in the states, the method comprising the steps of:

’925 Patent		
		generating, from the set of states of the first speech recognizer, a set of states of the second speech recognizer by selecting a subset of states of the first speech recognizer being distinctive of a particular application; and generating, from the set of probability density functions of the first speech recognizer, a set of probability density functions of the second speech recognizer by selecting a subset of probability density functions of the first speech recognizer being distinctive of the particular application, such that the second speech recognizer is at least one tailored to the particular application and requires reduced resources compared to the first speech recognizer. 2. The method of claim 1, further comprising the step of generating acoustic model parameters of the second speech recognizer by re-estimating acoustic model parameters of the first speech recognizer based on the set of states of the second speech recognizer and the set of probability density functions of the second speech recognizer. 4. The method of claim 2, wherein selecting at least one of the subset of states and the subset of probability density functions of the first speech recognizer exploits phonetical knowledge of the particular application. 5. The method of claim 4, wherein selecting at least one of the subset of states and the subset of probability density functions of the first speech recognizer exploits application specific training data. 6. The method of claim 5, **wherein selecting the subset of states comprises associating a multitude of speech frames of the training data with the correct states of the first speech recognizer and selecting those states with a frequency of occurrence above a threshold as the subset of states**.” Fischer, claim 6 (emphasis added).
Claim 14		

’925 Patent		
14.pre	A machine-readable storage medium, having stored thereon a computer program having a plurality of code sections executable by a machine for causing the machine to automatically generate from a first speech recognizer a second speech recognizer, said first speech recognizer comprising a first acoustic model with a first decision network and corresponding first phonetic contexts, and said second speech recognizer being adapted to a specific domain, said machine-readable storage causing the machine to perform the steps of:	See claim 1.pre (same reasoning from Fischer claim 6 but applying Fischer claim 15 or claim 6). *See also* Fischer Claim 15 (emphasis added). “10. **Apparatus for automatically generating, from a first Speech recognizer, a Second speech recognizer, wherein the first Speech recognizer includes a set of States and a set of probability density functions assembling output probabilities for an observation of a speech frame in the States, the apparatus comprising**: **at least one processor operative to**: (i) generate, from the Set of States of the first Speech recognizer, a set of States of the Second Speech recognizer by Selecting a Subset of States of the first speech recognizer **being distinctive of a particular application**; and (ii) generate, from the Set of probability density functions of the first speech recognizer, a set of probability density functions of the Second Speech recognizer by Selecting a Subset of probability density functions of the first Speech recognizer **being distinctive of the particular application**, Such that the Second speech recognizer is at least one of tailored to the particular application and requires reduced resources compared to the first Speech recognizer. 11. The apparatus of claim 10, wherein the at least one processor is further operative to generate acoustic model parameters of the Second Speech recognizer by re-estimating acoustic model parameters of the first Speech recognizer based on the Set of States of the Second Speech recognizer and the set of probability density functions of the second Speech recognizer. 12. The apparatus of claim 11, wherein the at least one processor is further operative to: (i) determine at least one of resource requirements and recognition accuracy of the Second speech recognizer; and (ii) repeat the State Set generation, probability density function Set generation, and acoustic model parameter generation operations, with one of more limiting and less limiting Selection criteria, when at least one of the resource requirements and the recognition accuracy does not achieve at least one of a resource target and an accuracy target, respectively.

’925 Patent		
		13. The apparatus of claim 11, wherein the operation of Selecting at least one of the Subset of States and the Subset of probability density functions of the first Speech recognizer exploits phonetical knowledge of the particular application. 14. The apparatus of claim 13, wherein the operation of Selecting at least one of the Subset of States and the Subset of probability density functions of the first Speech recognizer exploits application-specific training data. 15. The apparatus of claim 14, wherein the operation of Selecting the Subset of States comprises associating a multitude of Speech frames of the training data with the correct States of the first Speech recognizer and Selecting those States with a frequency.”
14.a	based on said first acoustic model, generating a second acoustic model with a second decision network and corresponding second phonetic contexts for said second speech recognizer by re-estimating said first decision network and said corresponding first phonetic contexts based on domain-specific training data, wherein said first decision network and said second decision network utilize a phonetic decision tree to perform speech recognition operations,	*See* claim 1.a (same reasoning from Fischer claim 6 but applying Fischer claim 15 or claim 6).and 14.pre. *See also* Fischer Claim 15 (emphasis added). “10. Apparatus for automatically generating, from a first Speech recognizer, a Second speech recognizer, wherein the first Speech recognizer includes a set of States and a set of probability density functions assembling output probabilities for an observation of a speech frame in the States, the apparatus comprising: at least one processor operative to: (i) **generate, from the Set of States of the first Speech recognizer, a set of States of the Second Speech recognizer by Selecting a Subset of States of the first speech recognizer** being distinctive of a particular application; and (ii) **generate, from the Set of probability density functions of the first speech recognizer, a set of probability density functions of the Second Speech recognizer by Selecting a Subset of probability density functions of the first Speech recognizer** being distinctive of the particular application, Such that the Second speech recognizer is at least one of tailored to the particular application and requires reduced resources compared to the first Speech recognizer. 11. The apparatus of claim 10, wherein the at least one processor is further operative to generate acoustic model parameters of the Second Speech recognizer by reestimating

’925 Patent		
		acoustic model parameters of the first Speech recognizer based on the Set of States of the Second Speech recognizer and the set of probability density functions of the second Speech recognizer. 12. The apparatus of claim 11, wherein the at least one processor is further operative to: (i) determine at least one of resource requirements and recognition accuracy of the Second speech recognizer; and (ii) repeat the State Set generation, probability density function Set generation, and acoustic model parameter generation operations, with one of more limiting and less limiting Selection criteria, when at least one of the resource requirements and the recognition accuracy does not achieve at least one of a resource target and an accuracy target, respectively. 13. The apparatus of claim 11, wherein the operation of Selecting at least one of the Subset of States and the Subset of probability density functions of the first Speech recognizer exploits phonetical knowledge of the particular application. 14. The apparatus of claim 13, **wherein the operation of Selecting at least one of the Subset of States and the Subset of probability density functions of the first Speech recognizer exploits application-specific training data**. 15. The apparatus of claim 14, wherein the operation of Selecting the Subset of States comprises associating a multitude of Speech frames of the training data with the correct States of the first Speech recognizer and Selecting those States with a frequency.” In addition, a POSITA would have found the addition of new phonetic contexts and nodes (which is not required by the claims) is an obvious variant to Fischer Claim 6. Moreover, this is obvious in light of Fischer alone or in combination with Sabourin or Schultz. A POSITA would have recognized that Sabourin and Fischer are in similar fields of invention. To the extent, re-estimation requires adding nodes, a POSITA would have looked to similar art in the field and recognized it was well known to add or delete nodes in the decision tree based on at least Sabourin and Schultz.

’925 Patent		
		Moreover, given this disclosure, the re-estimation of the first speech recognizer was obvious in light of Fischer alone, or in combination with Sabourin or Schultz. The motivation to combine these references would at least include: • Combining prior art elements according to known methods to yield predictable results (adding nodes to account for new phones in a domain or deleting unused phones were known using similar techniques for predictable results); • Simple substitution of one known element for another to obtain predictable results; • Use of known technique to improve similar devices (methods, or products) in the same way (same technique of adapting the decision network when expanding the domain of the recognizer to include new phones or deleting unused phones was known)); • Applying a known technique to a known device (method, or product) ready for improvement to yield predictable results (adding and deleting nodes in the decision network during adaptation were known using similar known techniques for predictable results); • “Obvious to try”—choosing from a finite number of identified, predictable solutions, with a reasonable expectation of success (obvious to add nodes for new sounds or remove nodes for unused phones); • Market forces and benefits associated with the known benefits of adding and deleting nodes were predictable to POSA • Teaching of prior art would have lead a POSA to combine the references to arrive at a adaptation of a recognizer that includes adding or deleting nodes. For example, Sabourin and Schultz explicitly re-estimate by adding nodes, pruning nodes and merging nodes in the decision network.

’925 Patent		
		It would be obvious to a person having ordinary skill in the art to combine Fischer with Sabourin and Schultz to disclose deleting nodes, adding nodes, pruning nodes and merging nodes in the decision network. Sabourin, Schultz, and Fischer are in the same field of art. A person having ordinary skill in the art would be motivated to combine Fischer with Sabourin or Schultz. A person having ordinary skill in the art considering Fischer’s disclosure that resources can be re-invested in to the speech recognition system would be motivated to seek out a system that trains the first decision network by adding nodes to account for new sounds, as disclosed by Sabourin and Schultz, which is a way of re-investing nodes. A person having ordinary skill in the art would have a reasonable expectation of success in implementing the teaching of Fischer to re-estimate the first decision network through addition of nods and deleting of nodes as provided by Sabourin and Schultz because the combination involves the predictable use of prior art elements according to their established functions. *See* Sabourin at Abstract, 2:57-3:6, 3:52-4:7, 4:16-45, 5:66-6:44, 6:45-7:8, 8:49-10:23; Schultz at Abstract, Section 3.3., Section 3.4.
14.b	wherein the number of nodes in the second decision network is not fixed by the number of nodes in the first decision network, and wherein said re-estimating comprises partitioning said training data using said first decision network of said first speech recognizer.	*See* claim 1.b (same reasoning from Fischer claim 6 but applying Fischer claim 15 or claim 6). *See also* Fischer Claim 15 (emphasis added). “10. Apparatus for automatically generating, from a first Speech recognizer, a Second speech recognizer, wherein the first Speech recognizer includes a set of States and a set of probability density functions assembling output probabilities for an observation of a speech frame in the States, the apparatus comprising: at least one processor operative to: (i) **generate, from the Set of States of the first Speech recognizer, a set of States of the Second Speech recognizer by Selecting a Subset of States of the first speech recognizer** being distinctive of a particular application; and (ii) **generate, from the Set of probability density functions of the**

’925 Patent		
		first speech recognizer, a set of probability density functions of the Second Speech recognizer by Selecting a Subset of probability density functions of the first Speech recognizer being distinctive of the particular application, Such that the Second speech recognizer is at least one of tailored to the particular application and requires reduced resources compared to the first Speech recognizer. 11. The apparatus of claim 10, wherein the at least one processor is further operative to generate acoustic model parameters of the Second Speech recognizer by reestimating acoustic model parameters of the first Speech recognizer based on the Set of States of the Second Speech recognizer and the set of probability density functions of the second Speech recognizer. 12. The apparatus of claim 11, wherein the at least one processor is further operative to: (i) determine at least one of resource requirements and recognition accuracy of the Second speech recognizer; and (ii) repeat the State Set generation, probability density function Set generation, and acoustic model parameter generation operations, with one of more limiting and less limiting Selection criteria, when at least one of the resource requirements and the recognition accuracy does not achieve at least one of a resource target and an accuracy target, respectively. 13. The apparatus of claim 11, wherein the operation of Selecting at least one of the Subset of States and the Subset of probability density functions of the first Speech recognizer exploits phonetical knowledge of the particular application. 14. The apparatus of claim 13, wherein the operation of Selecting at least one of the Subset of States and the Subset of probability density functions of the first Speech recognizer exploits application-specific training data. 15. The apparatus of claim 14, **wherein the operation of Selecting the Subset of States comprises associating a multitude of Speech frames of the training data with the correct States of the first Speech recognizer and Selecting those States with a frequency**.”

’925 Patent		
Claim 27		
27.pre	“A computerized method of generating a second speech recognizer comprising the steps of:	*See* claim 1, including 1.pre.a-c.
27.a	identifying a first speech recognizer of a first domain comprising a first acoustic model with a first decision network and corresponding first phonetic contexts;”	*See* claim 1, including 1.pre.b and 27.pre.
27.b	“receiving domain-specific training data of a second domain; and”	Fischer discloses receiving domain-specific training data of a second domain. *See* claim 1 including 1.pre.c and 1.a.1. *See also:* “1. A computer-based method of automatically generating, from a first speech recognizer, a second speech recognizer wherein the first speech recognizer includes a set of states and a set of probability functions assembling output probabilities for an observation of a speech frame in the states, the method comprising the steps of: generating, from the set of states of the first speech recognizer, a set of states of the second speech recognizer by selecting a subset of states of the first speech recognizer **being distinctive of a particular application**; and generating, from the set of probability density functions of the first speech recognizer, a set of probability density functions of the second speech recognizer by selecting a subset of probability density functions of the first speech recognizer **being distinctive of the particular application, such that the second speech recognizer is at least one tailored to the particular application** and requires reduced resources compared to the first speech recognizer.

’925 Patent		
		2. The method of claim 1, further comprising the step of generating acoustic model parameters of the second speech recognizer by re-estimating acoustic model parameters of the first speech recognizer based on the set of states of the second speech recognizer and the set of probability density functions of the second speech recognizer. 4. The method of claim 2, wherein selecting at least one of the subset of states and the subset of probability density functions of the first speech recognizer exploits phonetical knowledge of the particular application. 5. The method of claim 4, **wherein selecting at least one of the subset of states and the subset of probability density functions of the first speech recognizer exploits application specific training data**. 6. The method of claim 5, wherein selecting the subset of states comprises associating a multitude of speech frames of the training data with the correct states of the first speech recognizer and selecting those states with a frequency of occurrence above a threshold as the subset of states.” Fischer, claim 6 (emphasis added).
27.c	“based on the first speech recognizer and the domain-specific training data, generating a second acoustic model of said first domain and said second domain comprising a second acoustic model with a second decision network and corresponding second phonetic contexts, wherein the first domain comprises at least a first language, wherein the second domain comprises at least a	Fischer discloses based on the first speech recognizer and the domain-specific training data, generating a second acoustic model of said first domain and said second domain comprising a second acoustic model with a second decision network and corresponding second phonetic contexts. Claim 27 is an obvious variant of Fischer claim 6 because a POSITA would have found it obvious to add nodes to account for phones in the training data missing in the first speech recognizer. *See claim 1 including 1.pre.c and 1.a.1.* While Fischer does not explicitly contemplate that the final speech recognizer will be able to understand multiple languages, Fischer contemplates that its solution avoids the cost with adapting acoustic models for example, in connection with different applications within a dialect or application. Fischer further contemplates that once Fischer builds the speech recognition it can re-invest the saved resources into the system. *See* Fischer at 9:13-18. A POSITA would understand that this same adaptation could be

’925 Patent		
	second language, and wherein the second speech recognizer is a multi-lingual speech recognizer.”	performed with a pre-existing multi-lingual recognizer or used to create a multilingual recognizer. This was a common application and need for modified recognizers. For example, multilingual speech recognizers were well known at the time of the ’925 patent. *See, e.g.*, Schultz et al., “*Polyphone Decision Tree Specialization for Language Adaptation*”, ICASSP-2000, Istanbul, Turkey, Jun. 2000 (Section 2.2. describing generating and use of multilingual speech recognizers). A POSITA would have known that multilingual speech recognizers were well known and desirable. Given this disclosure, the creation of a multilingual second recognizer was obvious in light of Fischer alone, or in combination with Sabourin. The motivation to combine these references would at least include: • Combining prior art elements according to known methods to yield predictable results (multilingual recognizers were known using similar techniques for predictable results); • Simple substitution of one known element for another to obtain predictable results (substituting the domain from just one language to two); • Use of known technique to improve similar devices (methods, or products) in the same way (same technique of expanding domain of the recognizer to include multiple languages was known)); • Applying a known technique to a known device (method, or product) ready for improvement to yield predictable results (multilingual recognizers were known using similar known techniques for predictable results); • “Obvious to try”—choosing from a finite number of identified, predictable solutions, with a reasonable expectation of success (obvious to try to recognize more than one language);

’925 Patent		
		• Market forces and benefits associated with the known benefits of multilingual recognizers were predictable to POSA • Teaching of prior art would have lead a POSA to combine the references to arrive at a multilingual recognizer. For example, Sabourin explicitly discloses a multilingual system. Sabourin also discloses that there are two different language domains as recited. “The invention relates to a method and apparatus for training a multilingual speech model Set. The multilingual Speech model Set generated is Suitable for use by a speech recognition System for recognizing spoken utterances for at least two different languages. The invention allows using a single Speech recognition unit with a single Speech model Set to perform Speech recognition on utterances from two or more languages. The method and apparatus make use of a group of a group of acoustic Sub-word units comprised of a first Subgroup of acoustic Sub-word units associated to a first language and a Second Subgroup of acoustic Sub-word units associated to a Second language where the first Subgroup and the Second Subgroup share at least one common acoustic Sub-word unit. The method and apparatus also make use of a plurality of letter to acoustic Sub-word unit rules Sets, each letter to acoustic Sub-word unit rules Set being associated to a different language. A Set of untrained speech models is trained on the basis of a training Set comprising speech tokens and their associated labels in combination with the group of acoustic Sub-word units and the plurality of letter to acoustic Sub-word unit rules Sets. The invention also provides a computer readable Storage medium comprising a program element for implementing the method for training a multilingual Speech model Set.” Sabourin at Abstract. “A deficiency of the above-described method is that the Speech recognition System requires as an input the language associated to the input utterance, which may not be readily available to the Speech recognition System. Usually, obtaining the language requires prompting the user for the language of use thereby requiring an additionally

’925 Patent		
		Step in the Service being provided by the Speech recognition enabled system which may lower the level of satisfaction of the user with the system as a whole. Another deficiency of the above noted method is the costs associated to developing and maintaining a Speech model Set for each language the Speech recognition System is adapted to recognize. More Specifically, each speech model Set must be trained individually, a task requiring manpower for each individual language thereby increasing significantly the cost of Speech recognition Systems operating in multilingual environments with respect to Systems operating in unilingual environments. In addition, the above-described method requires the Storage of a speech model Set for each language in memory thereby increasing the cost of the Speech recognition System in terms of memory requirements. Finally, the above described method requires testing a speech model Set for each language thereby increasing the testing cost of the Speech recognition System for each language the Speech recognition System is adapted to recognize. Thus, there exists a need in the industry to refine the process of training Speech models So as to obtain an improved multilingual Speech model Set capable of being used by a speech recognition System for recognizing spoken utterances for at least two different languages.” Sabourin at 1:55-2:17. “In accordance with another broad aspect, the invention provides a method for generating a multilingual speech model Set Suitable for use in a multilingual speech recognition System. The method comprises providing a group of acoustic Sub-word units having a first Subgroup of acoustic Sub-word units associated to a first language and a Second Subgroup of acoustic Sub-word units associated to a Second language. The first Subgroup and the Second Subgroup share at least one common acoustic Sub-word unit. The method further comprises providing a training Set comprising a plurality of entries, each entry having a speech token representative of a word and a label being an orthographic representation of the word. The method further comprises providing a Set of untrained speech models and training the Set of untrained speech models by utilizing the training Set, the plurality of letter to acoustic Sub-word unit rules Sets and the group of acoustic Sub-word units to derive the multilingual Speech model Set.”

’925 Patent		
		Sabourin at 2:57-3:7. “The invention provides a method for generating the multilingual speech model set shown in FIG.1. As shown in FIG. 2 of the drawings, the method comprises providing 200 a group of acoustic Sub-word units comprised of a first Subgroup of acoustic Sub-word units associated to a first language and a Second Subgroup of acoustic Sub-word units associated to a Second language. The first Subgroup and the Second Subgroup share at least one common acoustic Sub word unit. In a preferred embodiment, the group of acoustic Sub word unit is obtained by combining sub-word units from individual languages. For each language for which the multilingual Speech model Set is to be representative of, an inventory of sub-word units can be obtained. The acoustic Sub-word units are basic units of Sound in a language. These inventories may be produced by phoneticians or may be provided as off-the-shelf components. In a specific example, the Sub-word units are phonemes. The Skilled person in the art will readily see that other sub-word units may be used here without detracting from the spirit of the invention.” Sabourin at 4:25-45. “The training of the set of untrained speech models further comprises processing 304 the group of transcriptions generated at step 300 on the basis of a speech token of the corresponding entry in the training Set whereby training the Set of untrained speech models to derive the multilingual Speech model Set.” Sabourin at 9:34-39. *See* Sabourin at Figs. 2-7. *See, e.g.*, Sabourin, Claim 1 (and other claims). It would be obvious to a person having ordinary skill in the art to combine Fischer with Sabourin. Both Fischer and Sabourin are in the same field of art. A person having

’925 Patent		
		ordinary skill in the art would be motivated to combine Fischer with Sabourin. A person having ordinary skill in the art considering Fischer discloses investing again resources into the speech recognizer and the existence of any missing nodes, would be motivated to seek out a system that builds a multilingual or different language speech set when the new domain is a new language, as disclosed by Sabourin. A person having ordinary skill in the art would have a reasonable expectation of success in implementing the teaching of Fischer to generate a multilingual or new language speech recognizer as provided by Sabourin because the combination at least involves the predictable use of prior art elements according to their established functions. Further, to the extent re-estimation requires also adding nodes to the second speech recognizer, Schultz all similarly disclose adding nodes. For the reasons above, a POSITA would similarly be motivated to combine either Schultz with Fischer to create a multilingual speech recognizer where re-estimation includes the addition of nodes to account for new phones or contexts not present in the first speech recognizer. • *See* Exhibit A-4, Claim 1.a.1.

APPENDIX B-1
Invalidity Claim Chart for U.S. Pat. No. 8,532,993 ('993 patent)
Mirjam Wester, *Pronunciation Variation Modeling for Dutch Automatic Speech Recognition*, 2002 ("Wester")[1]

On October 16, 2020, Nuance narrowed the asserted '993 patent claims to 17 and 19. Nuance's Disclosure Limiting Asserted Claims of Phase 1 Patents. Claims 17 and 19 of the '993 patent are anticipated and/or rendered obvious by Wester alone or in combination with at least the following references. The chart below discloses how prior art references identified by Omilia disclose, either expressly or inherently, and/or render obvious each of the asserted claims:

(1) Steinbiss et al., *The Philips research system for large-vocabulary continuous-speech recognition*, Proc. of 3rd European Conference on Speech Communication and Technology EUROSPEECH '93, 2125-2128 (1993) ("Steinbiss")[2]

(2) Jain, et al., *Creating Speaker-Specific Phonetic Templates with a Speaker-Independent Phonetic Recognizer: Implications for Voice Dialing*, IEEE International Conference on Acoustics, Speech, and Signal Processing Conference Proceedings, 881-884 (1996) ("Jain")[3]

Wester incorporates the following prior-art references:

- Kessens et al., *Improving the performance of a Dutch CSR by modeling within-word and cross-word pronunciation variation*, Speech Communication 29 (1999) 193-207 (incorporated as pp. 49-65 in Wester) ("Kessens")
- Wester, *Pronunciation Modeling for ASR – Knowledge-based and Data-derived Methods*, 2001 (incorporated as pp. 97-122 in Wester) ("Wester2")

This invalidity claim chart is based on Omilia's present understanding of the asserted claims, the Court's August 6, 2020 Claim Construction Order (ECF. No. 241), Plaintiff ("Nuance's") initial and all subsequent supplemental Infringement Contentions, its Response to Omilia NLS' Supplemental Preliminary Non-Infringement and Invalidity Contentions, its Response to Omilia NLS' Interrogatory No. 9, and Omilia's investigation to date. Omilia is not adopting Nuance's apparent constructions, nor is Omilia admitting to the accuracy of any particular construction. Omilia reserves all rights to amend this invalidity claim chart if Nuance further amends

[1] Wester was published in 2002. Wester is prior art under at least pre-AIA 35 U.S.C. § 102(a), and (b).
[2] Steinbiss was published in September, 1993. Steinbiss is prior art under at least pre-AIA 35 U.S.C. § 102(a), and (b).
[3] Jain was published in 1996. Jain is prior art under at least pre-AIA 35 U.S.C. § 102(a), and (b).

its Infringement Contentions, or in response to any statements made by Nuance at any phase of this litigation. In addition, Discovery is still ongoing, and Omilia reserves its right to supplement its contentions with additional evidence as discovery continues.

The citations provided are exemplary and do not necessarily include each and every disclosure of the limitation in the references. Omilia has cited the most relevant portions of the identified prior art. Other portions of the identified prior art may additionally disclose, either expressly or inherently, and/or render obvious one or more limitations of the asserted claims. To provide context or aid in understanding the prior art and the state of the art at the time of the alleged invention, Omilia reserves the right to rely on: (1) uncited portions of the identified prior art; (2) other prior art not identified herein; (3) references that show the state of the art (irrespective of whether such references themselves qualify as prior art to the asserted patents); (4) factual testimony from the inventors or authors of the prior art references, or purveyors of prior art devices; and/or (5) expert testimony.

Citations to a particular drawing or figure in the accompanying charts should be understood to include the description of the drawing or figure, as well as any text associated with the drawing or figure. Citations to particular text concerning a drawing or figure, should also be understood to encompass that drawing or figure as well. The identified prior art expressly and/or inherently discloses the features of the asserted claims under all proposed constructions of the asserted claims. To establish the inherency of certain features of the prior art to invalidate the asserted claims, Omilia may rely on cited or uncited portions of the prior art, other documents, factual testimony, and expert testimony. To the extent that the Court finds that this reference does not explicitly teach certain limitations in the asserted claims, such limitations would have been inherent and/or obvious and Omilia reserves the right to supplement or amend this claim chart to address such a finding.

’993 Patent		
Claim 17		
17.pre	A computer-readable storage device having instructions stored which, when executed on a processor, cause the processor to perform operations comprising:	To the extent that the preamble is limiting, Wester discloses, expressly or inherently, “computer-readable storage device having instructions stored which, when executed on a processor, cause the processor to perform operations:” In Wester, the process described was developed and implemented using continuous speech recognition computer system using conventional computer elements. *See, e.g.*, “Two continuous speech recognition (CSR) systems for Dutch are used in the publications in this thesis: the Phicos recognition system (Steinbiss et al. 1993), and the ICSI hybrid ANN/HMM speech recognition system (Bourlard and Morgan 1993). The main differences between the Phicos and ICSI systems are the search strategies that are used and the manner in which the acoustic probabilities are

’993 Patent		
		estimated. The ICSI system uses stack decoding and neural networks are employed to estimate the acoustic probabilities, whereas in the Phicos system, a Viterbi beam search is employed and mixtures of Gaussians are used to estimate the acoustic probabilities.” Wester, at 8 “In the Phicos recognition system (Steinbiss et al. 1993), continuous density hidden Markov models (HMMs) with 32 Gaussians per state are used. The HMMs have a tripartite structure, and each of the three parts consists of two states with identical emission distributions. The transition probabilities, which allow for loops, jumps and skips, are tied over all states. Feature extraction is carried out every 10 ms for 16 ms frames. The first step in the feature analysis is anFFT analysis to calculate the spectrum. In the following step, the energy in 14 mel-scaled filter bands between 350 and 3400 Hz is calculated. Next, a discrete cosine transformation is applied to the log of the filterband coefficients. The final processing stage is a running cepstral mean subtraction. In addition to the 14 cepstral coefficients, 14 delta coefficients are calculated, which makes a total of 28 feature coefficients, which are used to describe the speech signal.” Wester at 9. “The neural network in the ICSI hybrid HMM/ANN speech recognition system (Bourlard and Morgan 1993) was bootstrapped using segmentations of the training material obtained with the Phicos system. These segmentations were obtained by performing a Viterbi alignment using a baseline lexicon (only canonical pronunciations) and Phicos baseline acoustic models, i.e. no pronunciation variation had been explicitly modeled. The front-end acoustic processing consisted of calculating 12ift--order PLP features (Hermansky 1990), and energy every 10 ms, for 25 ms frames. The neural net takes an acoustic feature vector plus additional context from eight surrounding frames of features at the input, and outputs phone posterior probability estimates. The neural network has a hidden layer size of 1000 units and the same network was employed in all experiments.” Wester at 9.

’993 Patent		
		In addition, it would be obvious to a person having ordinary skill in the art to combine Wester with Steinbiss. Both Wester and Steinbiss are in the same field of art. Wester discloses that the speech recognition system of Steinbiss is “used in the publications in this thesis.” Wester, at 9. A person having ordinary skill in the art would therefore be motivated to use Steinbiss to implement the system of Wester. Steinbiss further teaches that “[s]peech recognition runs remotely on a PC which is connected to the network.” Steinbiss, at 2128. A person having ordinary skill in the art considering Wester would therefore understand that the system disclosed by Wester would be implemented on a computer, as taught by Steinbiss. A person having ordinary skill in the art would have a reasonable expectation of success in implementing Wester on the computer of Steinbiss because the combination involves the predictable use of prior art elements according to their established functions. The motivation to combine these references would at least include: • Combining prior art elements according to known methods to yield predictable results (using a computer to perform speech recognition using computers was well known in the art); • Simple substitution of one known element for another to obtain predictable results (implementing the system of Wester on top of the system of Steinbiss using a computer); • Use of known technique to improve similar devices (methods, or products) in the same way (the technique of using a computer to perform speech recognition was known); • Applying a known technique to a known device (method, or product) ready for improvement to yield predictable results (using a computer to perform speech recognition using computers was well known in the art); • “Obvious to try”—choosing from a finite number of identified, predictable solutions, with a reasonable expectation of success (Wester teaches that Steinbiss may be used as a base system);

’993 Patent		
		• Market forces and benefits associated with the known benefits of automatic speech recognition Teaching of prior art would have lead a POSA to combine the references to arrive at a speech recognition system implemented on a computer. Steinbiss also discloses, expressly or inherently, “computer-readable storage device having instructions stored which, when executed on a processor, cause the processor to perform operations:” In Steinbiss, the process described was developed and implemented using PC based implementation. *See, e.g.*, “The system has been successfully applied to the American English DARPA RM task. Here, we report experimental results for a German 13 000-word Philips internal dictation task. In addition to the scientific prototype, a PC version has been set up which is described here for the first time.” Steinbiss, at 2125. “The organization of the paper is as follows. We first summarize the statistical approach to speech recognition and then describe the four main entities of our system: acoustic analysis, acousticphonetic modelling, language modelling and search. A section with experiments on our internal dictation task follows. The final section describes a PC based implementation of our system.” Steinbiss, at 2125. “8. A PC Based Continuous Speech-Recognition System for Dictation . . . The system developed at Philips Dictation Systems, Vienna, and described here adopts this non-interactive approach and thus allows the person to dictate with a natural speaking style. After the

’993 Patent		
		speech is processed by the speech recognizer, the secretary has only to correct the recognition errors, which is both faster and a more interesting job to do.” Steinbiss, at 2128. “Speech recognition runs remotely on a PC which is connected to the network. An acoustic front-end performs the acoustic analysis. Recognition is sped up by a dedicated co-processor board containing application-specific ICs. Depending on the speaker and the specific boundary conditions, recognition with a 10 - 20 000-word vocabulary runs in 1 - 3 times real-time.” Steinbiss, at 2128.
17.a	approximating transcribed speech using a phonemic transcription dataset associated with a speaker, to yield a language model, where the phonemic transcription dataset is based on a pronunciation model of the speaker;	Wester discloses, expressly or inherently, the step of “approximating transcribed speech using a phonemic transcription dataset associated with a speaker, to yield a language model, where the phonemic transcription dataset is based on a pronunciation model of the speaker:” Wester discloses, expressly or inherently, the step of “***approximating transcribed speech using a phonemic transcription dataset associated with a speaker***, to yield a language model, where the phonemic transcription dataset is based on a pronunciation model of the speaker:” In Wester, a speech recognition system is trained using phonetic transcriptions of training material. Wester further discloses the training data may include “rate of speech or type of dialect.” A POSA would further understand that this training data is “associated with a speaker.” *See, e.g.*, “**Lexicon.** The lexicon (or dictionary as it is often referred to) typically consists of the orthography of words that occur in the training material and their corresponding phonetic transcriptions. During recognition, the phonetic transcriptions in the lexicon function as a constraint which defines the sequences of phonemes that are permitted to occur. The transcriptions can be obtained either manually or through grapheme-to-phoneme conversion. In pronunciation variation research one is usually confronted with two types of lexica: a canonical (or baseline) lexicon and a multiple pronunciation lexicon. A canonical lexicon contains the normative or standard transcriptions for the words; this is a single transcription per word. A

'993 Patent		
		multiple pronunciation lexicon contains more than one variant per word, for some or all of the words in the lexicon." Wester, at 7. "Within the search strategy, a single-pass or multi-pass search can be employed. In the work presented in this thesis, only single-pass search strategies have been employed. However, it has been shown that multi-pass searches can be very useful for pronunciation modeling, as this makes it possible to dynamically change the lexicon. Factors such as rate of speech or type of dialect, which are measured or estimated in a first pass, can be used to determine the appropriate set of pronunciations to include in the lexicon. This dynamically adjusted lexicon can then be employed in a second pass. Examples of pronunciation variation research in which a multi-pass approach has been used are Fosler-Lussier (1999) and Lee and Wellekens (2001)." Wester, at 7-8. In addition, to the extent that Wester does not incorporate Wester2 by reference, it would be obvious to a person having ordinary skill in the art to combine Wester with Wester2. Both Wester and Wester2 are in the same field of art. Wester discloses that it and Wester2 are part of the same research and that Wester2 provides results relied upon by Wester. Wester at 32. A person having ordinary skill in the art would therefore be motivated to use the disclosures of Wester2 to implement the system of Wester. A person having ordinary skill in the art would have a reasonable expectation of success in implementing Wester using the disclosure of Wester2 because the combination involves the predictable use of prior art elements according to their established functions. The motivation to combine these references would at least include: • Combining prior art elements according to known methods to yield predictable results (creating a language model for speech recognition as well known in the art);

’993 Patent		
		• Simple substitution of one known element for another to obtain predictable results (implementing the system of Wester2 as part of the system of Wester); • Use of known technique to improve similar devices (methods, or products) in the same way (the technique of creating language models for speech recognition was known); • Applying a known technique to a known device (method, or product) ready for improvement to yield predictable results (using data-driven techniques to create a language model for speech recognition was well known in the art); • “Obvious to try”—choosing from a finite number of identified, predictable solutions, with a reasonable expectation of success (Wester teaches that Wester2 forms the basis of the system); • Market forces and benefits associated with the known benefits of automatic speech recognition; and • Teaching of prior art would have lead a POSA to combine the references to arrive at a language model for speech recognition. Wester2 also discloses, expressly or inherently, the step of “***approximating transcribed speech using a phonemic transcription dataset associated with a speaker***, to yield a language model, where the phonemic transcription dataset is based on a pronunciation model of the speaker:” In Wester2, a speech recognition system is trained using phonetic transcriptions of training material. Wester2 further discloses the training data includes phonetic transcription of data that include pronunciation variation. A POSA would further understand that this training data is “associated with a speaker.” *See, e.g.*, “It is widely assumed that pronunciation variation is one of the factors which leads to less than optimal performance in automatic speech recognition (ASR) systems. Therefore, in the last few decades, effort has been put into finding solutions to deal with the difficulties linked to pronunciation variation. “Pronunciation variation” as a term could be used to describe most of the variation present in speech. The task of modeling it could consequently be seen as the task of

’993 Patent		
		solving the problem of ASR. However, this article has no pretension of going quite that far, seeing as we are not dealing with the full scope of pronunciation variation, but have restricted ourselves to pronunciation variation that becomes apparent in a careful broad phonetic (phonemic) transcription of the speech, in the form of insertions, deletions or substitutions of phones relative to the canonical transcription of the words. This type of pronunciation variation can be said to occur at the segmental level.” Wester2, at 99. In addition, to the extent that Wester does not incorporate Kessens by reference, it would be obvious to a person having ordinary skill in the art to combine Wester with Kessens. Both Wester and Kessens are in the same field of art. Wester discloses that it and Kessens are part of the same research and that Kessens provides results relied upon by Wester. Wester at 32. A person having ordinary skill in the art would therefore be motivated to use the disclosures of Kessens to implement the system of Wester. A person having ordinary skill in the art would have a reasonable expectation of success in implementing Wester using the disclosure of Kessens because the combination involves the predictable use of prior art elements according to their established functions. The motivation to combine these references would at least include: • Combining prior art elements according to known methods to yield predictable results (creating a language model for speech recognition as well known in the art); • Simple substitution of one known element for another to obtain predictable results (implementing the system of Wester2 as part of the system of Wester); • Use of known technique to improve similar devices (methods, or products) in the same way (the technique of creating language models for speech recognition was known);

’993 Patent		
		• Applying a known technique to a known device (method, or product) ready for improvement to yield predictable results (using data-driven techniques to create a language model for speech recognition was well known in the art); • “Obvious to try”—choosing from a finite number of identified, predictable solutions, with a reasonable expectation of success (Wester teaches that Wester2 forms the basis of the system); • Market forces and benefits associated with the known benefits of automatic speech recognition; and • Teaching of prior art would have lead a POSA to combine the references to arrive at a language model for speech recognition. Kessens also discloses, expressly or inherently, the step of “***approximating transcribed speech using a phonemic transcription dataset associated with a speaker***, to yield a language model, where the phonemic transcription dataset is based on a pronunciation model of the speaker:” In Kessens, a speech recognition system is trained using phonetic transcriptions of training material. Kessens further discloses the training data includes pronunciation variants. A POSA would further understand that this training data is “associated with a speaker.” *See, e.g.*, “In the third step, the language model is altered. To calculate the baseline language model the orthographic representation of the words in the training corpus is used. Because there is only one variant per word this suffices. However, when a multiple pronunciation lexicon is used during recognition and the language model is trained on the orthographic representation of the words, all variants of the same word will have equal a priori probabilities (this probability is determined by the language model). A drawback of this is that a sporadically occurring variant may have a high a priori probability because it is a variant of a frequently occurring word, whereas the variant should have a lower a priori probability on the basis of its occurrence. Consequently, the variant may be easily confused with other words in the lexicon. A way of reducing this confusability is to base the calculation of the language model on the phone transcription of the words instead of on the orthographic transcription, i.e. on the basis of the phone transcriptions of the corpus obtained

’993 Patent		
		through forced recognition. A recognition test is performed using this language model, the multiple pronunciation lexicon and the updated phone models (test condition: MMM).” Kessens, at 197. Wester discloses, expressly or inherently, the step of “approximating transcribed speech using a phonemic transcription dataset associated with a speaker, ***to yield a language model***, where the phonemic transcription dataset is based on a pronunciation model of the speaker:” In Wester, a language model is created using the training data, which includes pronunciation variation. *See, e.g.*, “**Language Model.** Typical recognizers use n-gram language models. An n-gram contains the prior probability of the occurrence of a word (unigram), or of a sequence of words (bigram, trigram etc.): unigram probability $P(w_i)$ (1.4) and bigram probability $P(w_i \| w_{i-1})$ (1.5) The prior probabilities (priors) in a language model are often estimated from large amounts of training texts for which there is no corresponding acoustic material, i.e., the training texts consist of text material *only*. In the studies presented in this thesis, this is not the case, as the training material used to train the acoustic models is also employed to estimate the probabilities in the language model (see Section 1.5 for more information on this speech material). This makes it possible to incorporate pronunciation variation in the language models, by estimating prior probabilities for the variants in the training corpus, rather than for the words.” Wester, at 7. “The systems use word-based unigram and bigram language models. The lexicon is the same in both systems, in the sense that it contains the orthography of the words and phone transcriptions for the pronunciations. However, it differs in the sense that the ICSI lexicon also contains prior probabilities for the variants of the words, whereas the Phicos lexicon does not. In the ICSI lexicon

’993 Patent		
		the prior probabilities are distributed over all variants for a word and add up to 1.0 for each word. Depending on the type of variants (knowledge-based or data-derived) the prior probabilities are distributed either equally over the variants of a word or they differ for the variants of a word as they are estimated from the training data.” Wester, at 9. “The main goal of the research presented in this thesis was to improve the performance of Dutch ASR. Statistically significant improvements in WER were found, both for the knowledge-based and data-derived approaches (Kessens et al. 1999a; Wester 2001). The results presented in publication 1 and 3 show that in order to obtain significant improvements in WERs, prior probabilities for the variants should be incorporated in the recognition process in addition to adding variants to the lexicon.” Wester, at 32. “In publication 1, another of our objectives was formulated as follows: ‘Our long-term goal is to find the set of rules which is optimal for modeling pronunciation variation.’ It is difficult to conclude whether this goal has been reached or not. It is possible that in the course of the research carried out for this thesis the optimal set of variants for the VIOS data was found. However, if that is the case, it went unnoticed, as we implicitly assumed that performing recognition with the optimal set of variants would lead to lower WERs. In Section 1.7.1, I argued that the reason for the lack of improvement in WER is because conditional probabilities are not taken into account in a static lexicon. Therefore, it could be the case that we have the correct set of variants to describe the pronunciation variation present in the VIOS material, but that this is not reflected in the WERs because of lexical confusability.” Wester, at 32. Wester discloses, expressly or inherently, the step of “approximating transcribed speech using a phonemic transcription dataset associated with a speaker, to yield a language model, ***where the phonemic transcription dataset is based on a pronunciation model of the speaker***:” In Wester, a

’993 Patent		
		speech recognition system is trained using phonetic transcriptions of training material. Wester further discloses the training data may include “rate of speech or type of dialect.” A POSA would further understand that this training data is “based on a pronunciation model of the speaker.” *See, e.g.*, “Information about pronunciation variation can be acquired from the data itself or through (prior) knowledge; also termed the data-derived and the knowledge-based approaches to modeling pronunciation variation. One can classify approaches in which information is derived from phonological or phonetic knowledge and/or linguistic literature (Cohen 1989; Giachin et al. 1991) under knowledge-based approaches. Existing dictionaries also fit into this category (Lamel and Adda 1996; Roach and Arnfield 1998). In contrast, data-derived approaches include methods in which manual transcriptions of the training data are employed to obtain information (Riley et al. 1999; Saraclar et al. 2000), or automatic transcriptions are used as the starting point for generating lists of variants (Fosler-Lussier 1999; Wester and Fosler-Lussier 2000).” Wester, at 13. “The options for incorporating the information into the ASR system are determined by the manner in which the variants are obtained. Using theoretical phonological rules limits the possibilities one has to merely adding variants, whereas a manual or good quality automatic transcription allows for more options. In the studies presented in this thesis both major approaches to obtaining variants have been used. In Kessens, Wester, and Strik (1999a) (publication 1), a knowledge based approach to obtaining pronunciation variants for Dutch is investigated. In Wester (2001) (publication 3), in addition to the knowledge-based approach, a data-derived approach is studied. In this study, a comparison is also made between the two approaches by analyzing the degree of overlap between the different lexica they produce.” Wester, at 13.

’993 Patent		
		In addition, as previously discussed, to the extent that Wester does not incorporate Wester2 by reference, it would be obvious to a person having ordinary skill in the art to combine Wester with Wester2. Wester2 also discloses, expressly or inherently, the step of “approximating transcribed speech using a phonemic transcription dataset associated with a speaker, to yield a language model, ***where the phonemic transcription dataset is based on a pronunciation model of the speaker***:” In Wester2, a speech recognition system is trained using phonetic transcriptions of training material. Wester2 further discloses the training data includes phonetic transcription of data that include pronunciation variation. A POSA would further understand that this training data is “based on a pronunciation model of the speaker.” *See, e.g.*, “Thus, it seems that the problem of modeling pronunciation variation lies in accurately predicting the word pronunciations that occur in the test material. In order to achieve this, the pronunciation variants must first be obtained in some way or other. Approaches that have been taken to modeling pronunciation variation can be roughly divided into pronunciation variants derived from a corpus of pronunciation data or from pre-specified phonological rules based on linguistic knowledge (Strik and Cucchiarini 1999). Both have their pros and cons. For instance, the information from linguistic literature is not exhaustive; many processes that occur in real speech are yet to be described. On the other hand, the problem with an approach that employs data to access information is that it is extremely difficult to extract *reliable* information from the data.” Wester2, at 100. “In the following section, the speech material is described. This is followed by a description of the standard set-up of the two recognition systems: the ICSI hybrid ANN/HMM speech recognition system (Bourlard and Morgan 1993) and the Phicos recognition system (Steinbiss et al. 1993). In section 3, the baseline results of the two systems are presented. Next, a description is given of how the various lexica pertaining to pronunciation modeling are created: the knowledge-based approach to generating new pronunciations and the data-derived approach to pronunciation modeling. In Section 5, an extended description of the confusability metric, proposed in Wester and Fosler-Lussier (2000), is given. This is followed by the results of recognition experiments employing the

’993 Patent		
		different pronunciation lexica. In section 7, comparisons are made as to which variants overlap in the different lexica. Finally, we end by discussing the implications of our results and shortly summarizing the most important findings of this research.” Wester2, at 101-02. It would be obvious to a person having ordinary skill in the art to combine Wester with Jain. Both Wester and Jain are in the same field of art. A person having ordinary skill in the art would be motivated to combine Wester with Jain. A person having ordinary skill in the art considering Wester’s disclosure that “pronunciation variation is one of the factors which leads to less than optimal performance in automatic speech recognition (ASR) systems” and “automatic transcriptions are used as the starting point for generating lists of variants” would be motivated to seek out a system that automatically captures a phonetic representation of pronunciation variants, as disclosed by Jain. A person having ordinary skill in the art would have a reasonable expectation of success in implementing Wester using the phonetic transcriptions provided by Jain because the combination involves the predictable use of prior art elements according to their established functions. The motivation to combine these references would at least include: • Combining prior art elements according to known methods to yield predictable results (using speaker-dependent training data to improve accuracy for a user); • Simple substitution of one known element for another to obtain predictable results (adding additional pronunciation variants to improve accuracy of a speech recognition system); • Use of known technique to improve similar devices (methods, or products) in the same way (using speaker-dependent pronunciation data to improve the accuracy of a speech recognition system);

’993 Patent		
		• Applying a known technique to a known device (method, or product) ready for improvement to yield predictable results (using speaker-dependent pronunciation was well known in the art); • “Obvious to try”—choosing from a finite number of identified, predictable solutions, with a reasonable expectation of success (there are a finite number of known techniques to improve accuracy for speech recognition systems); • Market forces and benefits associated with the known benefits of automatic speech recognition; and • Teaching of prior art would have lead a POSA to combine the references to arrive at a system that includes transcription data reflecting pronunciation data from a speaker. Jain also discloses, expressly or inherently, the step of “***approximating transcribed speech using a phonemic transcription dataset associated with a speaker***, to yield a language model, ***where the phonemic transcription dataset is based on a pronunciation model of the speaker***:” In Jain, speaker-specific models are used to transcribe telephone speech into its phonetic subparts. A POSA would further understand that the “telephone speech” is associated with a particular speaker. *See, e.g.*, “In this section, we describe the method for generating the speaker-specific models. Our speaker-specific models represent utterances as sequences of phonemes rather than acoustic parameters. A speaker-independent phonetic front end is used to generate the string of phonemes. For this approach to work, all that is needed is that the phonetic frontend behave in a consistent manner, i.e. it must produce the same (or nearly the same) string of phonemes each time the word is spoken.” Jain, at 881. “The following steps produce a phonetic transcription: 1. Data Capture: Telephone speech is sampled at 8 kHz. Unnecessary silence at the beginning and end of the utterance is removed by end-point detection.

<table>
<tr><th colspan="3"><u>’993 Patent</u></th></tr>
<tr><td></td><td></td><td>2. Feature Extraction: Seventh order RASTA features [2] are computed for every 10ms of speech using a 10ms window. This yields eight coefficients per frame.”

Jain, at 881.

“Another method is to force-align each phoneme string (which was generated using the template generation technique) with all the other waveforms corresponding to the same label and compute the average Viterbi score. The template with the maximum average score is selected as the word-model for that label. Table 1 shows the error rates obtained with forced-alignment. It can be seen that the error rates decreased substantially for all the 4 channels.

<table><tr><th>Telephone Channel</th><th>1</th><th>2</th><th>3</th><th>4</th></tr><tr><td>Baseline</td><td>7.9</td><td>12.0</td><td>8.7</td><td>18.5</td></tr><tr><td>Forced Alignment</td><td>4.8</td><td>8.0</td><td>5.8</td><td>15.5</td></tr></table>
Table 1. Effect of Forced Alignment ”

Jain, at 882-83.

“It is well known that speaker dependent systems perform better than speaker-independent systems. In several commercial products, speaker adaptation is used to adapt speaker-independent models to the current user.”

Jain at 883.</td></tr>
<tr><td>17.b</td><td>incorporating, into the language model, pronunciation probabilities associated with respective unique labels for each different</td><td>Wester discloses, expressly or inherently, the step of “incorporating, into the language model, pronunciation probabilities associated with respective unique labels for each different pronunciation of a word, wherein the respective unique label for a most frequent word indicates a special status in the language model.”

Wester discloses, expressly or inherently, the step of “incorporating, into the language model, pronunciation probabilities associated with respective unique labels for each different pronunciation of a word, wherein the respective unique label for a most frequent word indicates a</td></tr>
</table>

’993 Patent		
	pronunciation of a word, wherein the respective unique label for a most frequent word indicates a special status in the language model; and	special status in the language model.” In Wester, multiple pronunciations contain more than one variant for words and their probabilities, which are associated with unique labels. The variants and their probabilities are then incorporated into a language model. *See, e.g.*, “**Lexicon.** The lexicon (or dictionary as it is often referred to) typically consists of the orthography of words that occur in the training material and their corresponding phonetic transcriptions. During recognition, the phonetic transcriptions in the lexicon function as a constraint which defines the sequences of phonemes that are permitted to occur. The transcriptions can be obtained either manually or through grapheme-to-phoneme conversion. In pronunciation variation research one is usually confronted with two types of lexica: a canonical (or baseline) lexicon and a multiple pronunciation lexicon. A canonical lexicon contains the normative or standard transcriptions for the words; this is a single transcription per word. A multiple pronunciation lexicon contains more than one variant per word, for some or all of the words in the lexicon.” Wester, at 7. “**Language Model.** Typical recognizers use n-gram language models. An n-gram contains the prior probability of the occurrence of a word (unigram), or of a sequence of words (bigram, trigram etc.): unigram probability $P(w_i)$ (1.4) and bigram probability $P(w_i \mid w_{i-1})$ (1.5) The prior probabilities (priors) in a language model are often estimated from large amounts of training texts for which there is no corresponding acoustic material, i.e., the training texts consist of text material *only*. In the studies presented in this thesis, this is not the case, as the training material used to train the acoustic models is also employed to estimate the probabilities in the language model (see Section 1.5 for more information on this speech material). This makes it

’993 Patent		
		possible to incorporate pronunciation variation in the language models, by estimating prior probabilities for the variants in the training corpus, rather than for the words.” Wester, at 7. “Within the search strategy, a single-pass or multi-pass search can be employed. In the work presented in this thesis, only single-pass search strategies have been employed. However, it has been shown that multi-pass searches can be very useful for pronunciation modeling, as this makes it possible to dynamically change the lexicon. Factors such as rate of speech or type of dialect, which are measured or estimated in a first pass, can be used to determine the appropriate set of pronunciations to include in the lexicon. This dynamically adjusted lexicon can then be employed in a second pass. Examples of pronunciation variation research in which a multi-pass approach has been used are Fosler-Lussier (1999) and Lee and Wellekens (2001).” Wester, at 7-8. “The systems use word-based unigram and bigram language models. The lexicon is the same in both systems, in the sense that it contains the orthography of the words and phone transcriptions for the pronunciations. However, it differs in the sense that the ICSI lexicon also contains prior probabilities for the variants of the words, whereas the Phicos lexicon does not. In the ICSI lexicon the prior probabilities are distributed over all variants for a word and add up to 1.0 for each word. Depending on the type of variants (knowledge-based or dataderived2) the prior probabilities are distributed either equally over the variants of a word or they differ for the variants of a word as they are estimated from the training data.” Wester, at 9. “Information about pronunciation variation can be acquired from the data itself or through (prior) knowledge; also termed the data-derived and the knowledge-based approaches to modeling pronunciation variation. One can classify approaches in which information is derived from phonological or phonetic knowledge and/or linguistic literature (Cohen 1989; Giachin et al. 1991) under knowledge-based approaches. Existing dictionaries also fit into this category (Lamel and

’993 Patent		
		Adda 1996; Roach and Arnfield 1998). In contrast, data-derived approaches include methods in which manual transcriptions of the training data are employed to obtain information (Riley et al. 1999; Saraclar et al. 2000), or automatic transcriptions are used as the starting point for generating lists of variants (Fosler-Lussier 1999; Wester and Fosler-Lussier 2000).” Wester, at 13. “The options for incorporating the information into the ASR system are determined by the manner in which the variants are obtained. Using theoretical phonological rules limits the possibilities one has to merely adding variants, whereas a manual or good quality automatic transcription allows for more options. In the studies presented in this thesis both major approaches to obtaining variants have been used. In Kessens, Wester, and Strik (1999a) (publication 1), a knowledge based approach to obtaining pronunciation variants for Dutch is investigated. In Wester (2001) (publication 3), in addition to the knowledge-based approach, a data-derived approach is studied. In this study, a comparison is also made between the two approaches by analyzing the degree of overlap between the different lexica they produce.” Wester, at 13. “After the pronunciation variants are obtained, the next question that must be addressed is how the information should be incorporated into the ASR system. There are different levels at which this problem can be addressed. In Strik and Cucchiarini (1999) a distinction was made among incorporating information on pronunciation variation in the lexicon, the acoustic models and the language models. In the following sections, pronunciation modeling at each of these levels is discussed. First, adding variants to the lexicon is addressed. This is followed by a discussion of lexical confusability, which is an issue that is closely linked to modeling pronunciation variation in the lexicon. Next, the role of forced alignment in pronunciation modeling is explained, before discussing how pronunciation variation can be incorporated in the acoustic models and how the language models are employed in pronunciation modeling. The final issue that is addressed in this section is the use of articulatory-acoustic features in pronunciation modeling.”

’993 Patent		
		Wester, at 13-14. “As speech recognizers make use of a lexicon, pronunciation variation is often modeled at the level of the lexicon. Variation that occurs within a word can be dealt with in the lexicon by adding variants of the words to the lexicon. Variants of a single word are different phonetic transcriptions of one and the same word; i.e., substitutions, insertions and deletions of phones in relation to the base-form variant. This type of variation is within-word variation. However, in continuous speech a lot of variation occurs over word boundaries. This is referred to as cross-word variation. Cross-word variation can, to a certain extent, be dealt with in the lexicon by adding sequences of words which are treated as one entity, i.e., multi-words. The variation in pronunciation that occurs due to cross-word variation is modeled by adding variants of the multi-words to the lexicon (Sloboda and Waibel 1996; Fosler-Lussier and Williams 1999). An alternative method for modeling cross-word variation in the lexicon is described in Cremelie and Martens (1999): the cross-word variants are coded in the lexicon in such a way that during recognition only compatible variants can be interconnected. The importance of cross-word variation modeling was illustrated in Yang and Martens (2000) (the follow-up study to Cremelie and Martens (1999)) which shows that almost all the gain (relative improvement of 45% in WER over baseline performance) in their method is due to modeling cross-word variation.” Wester, at 14. “Incorporating pronunciation variation in the language model can be carried out by estimating the probabilities of the variants instead of the probabilities of the words. This is of course only possible if the pronunciation variants are transcribed in the training material, and the language models are trained on this material. An intermediate level of modeling pronunciation variation in the language model is possible in the form of word classes. In particular, this approach is taken to deal with processes of cross-word variation such as liaisons in French (Brieussel-Pousse and Perennou 1999).” Wester at 16.

’993 Patent		
		“Many studies (Cohen 1989; Yang and Martens 2000; Ma et al. 1998; Fosler-Lussier 1999) have shown that probabilities of the variants (or probabilities of rules) play an important role in whether an approach to modeling pronunciation variation is successful or not. Prior probabilities of the variants can be incorporated in the language model or in the lexicon, depending on the type of recognizer that is being used.” Wester at 17. “Incorporating variants in the language model is an integral part of the method for modeling pronunciation variation reported in (Kessens, Wester, and Strik 1999a) (publication 1). This approach is necessary as in the Phicos recognition system incorporating priors for the variants in the system is only possible through the language model. Incorporating priors for variants in the ICSI system is possible in the lexicon, thus obviating the need for priors of variants in the language model. Experiments investigating the effect of including or excluding priors during recognition are reported in Wester (2001) (publication 3).” Wester at 17. “The main goal of the research presented in this thesis was to improve the performance of Dutch ASR. Statistically significant improvements in WER were found, both for the knowledge-based and data-derived approaches (Kessens et al. 1999a; Wester 2001). The results presented in publication 1 and 3 show that in order to obtain significant improvements in WERs, prior probabilities for the variants should be incorporated in the recognition process in addition to adding variants to the lexicon.” Wester, at 32. “In publication 1, another of our objectives was formulated as follows: ‘Our long-term goal is to find the set of rules which is optimal for modeling pronunciation variation.’ It is difficult to conclude whether this goal has been reached or not. It is possible that in the course of the research carried out for this thesis the optimal set of variants for the VIOS data was found. However, if that is the case, it went unnoticed, as we implicitly assumed that performing recognition with the

’993 Patent		
		optimal set of variants would lead to lower WERs. In Section 1.7.1, I argued that the reason for the lack of improvement in WER is because conditional probabilities are not taken into account in a static lexicon. Therefore, it could be the case that we have the correct set of variants to describe the pronunciation variation present in the VIOS material, but that this is not reflected in the WERs because of lexical confusability.” Wester, at 32. “In Section 1.7, lexical confusability, phone transcriptions, and the beads-on-a-string paradigm were presented as shortcomings of the segmental approach to modeling pronunciation variation. This may give the impression that there is no future for pronunciation modeling. However, the outlook for pronunciation modeling is not quite that bleak. It is my impression that the future of pronunciation modeling should lie in employing different levels of linguistic information to predict and model the variation present in the speech material. This section gives a few examples of how this can be achieved in pronunciation modeling.” Wester at 32-33. “The results presented in publication 3 of this thesis show that simply adding a great deal of variants to the lexicon leads to a deterioration in WER. Therefore, prior probabilities are included in the decoding process. In Section 1.7.1, it was argued that although prior probabilities are important to include in the recognition process they do not suffice for modeling pronunciation variation and that conditional probabilities are possibly the key to reducing WERs.” Wester, at 33. In addition, to the extent that Wester does not incorporate Kessens by reference, as previously discussed with respect to limitation 17.a, which is hereby incorporated by reference, it would be obvious to a person having ordinary skill in the art to combine Wester with Kessens. Kessens also discloses, expressly or inherently, the step of “***incorporating, into the language model, pronunciation probabilities associated with respective unique labels for each different***

’993 Patent		
		pronunciation of a word, wherein the respective unique label for a most frequent word indicates a special status in the language model:” In Kessens, pronunciation variations are modeled with their associated probabilities, which are associated with unique labels. The pronunciation variants and their probabilities are then incorporated into a language model. *See, e.g.*, “The present research concerns the continuous speech recognition component of a spoken dialog system called OVIS (Strik et al., 1997). OVIS is employed to automate part of an existing Dutch public transport information service. A large number of telephone calls of the on-line version of OVIS have been recorded and are stored in a database called VIOS. The speech material consists of interactions between man and machine. The data clearly show that the manner in which people speak to OVIS varies, ranging from using hypoarticulated speech to hyper-articulated speech. As pronunciation variation degrades the performance of a continuous speech recognizer (CSR) - if it is not properly accounted for - solutions must be found to deal with this problem. We expect that by explicitly modeling pronunciation variation some of the errors introduced by the various ways in which people address the system will be corrected. Hence, our ultimate aim is to develop a method for modeling Dutch pronunciation variation which can be used to tackle the problem of pronunciation variation for Dutch CSRs.” Kessens, at 194. “Our experiments showed that modeling within-word pronunciation variation in the lexicon improves the CSR’s performance. However, in continuous speech there is also a lot of variation which occurs over word boundaries. For modeling crossword variation, various methods have been tested in the past (see e.g. Cremelie and Martens, 1998; Perennou and Brieussel-Pousse, 1998; Wiseman and Downey, 1998). In our previous research (Kessens and Wester, 1997), we showed that adding multi-words (i.e. sequences of words) and their variants to the lexicon can be beneficial. Therefore, we decided to retain this approach in the current research. However, we also tested a second method for modeling cross-word variation. For this method, we selected from the multi-words the set of words which are sensitive to the cross-word processes that we focus on; cliticization, reduction and contraction (Booij, 1995). Next, the variants of these words are added

’993 Patent		
		to the lexicon. In other words, in this approach no multi-words (or their variants) are added to the lexicon.” Kessens, at 195. “In this paper, we propose a general procedure for modeling pronunciation variation. This procedure affects all three levels of the CSR at which modeling can take place: i.e. the lexicon, the phone models and the language models (Strik and Cucchiarini, 1998). Table 1 shows at which levels pronunciation variation can be incorporated in the recognition process, and the different test conditions which are used to measure the effect of adding pronunciation variation. In the abbreviations used in Table 1, the first letter indicates which type of recognition lexicon was used; either a lexicon with single (S) or multiple (M) pronunciations per word. The second letter indicates whether single (S) or multiple (M) pronunciations per word were present in the corpus used for training the phone models. The third letter indicates whether the language model was based on words (S) or on the pronunciation variants of the words (M).” Kessens, at 195.

Table 1
The test conditions used to measure the effect modeling pronunciation variation

	Test condition	Lexicon	Phone models	Language models
Baseline	SSS	S	S	S
1	MSS	M	S	S
2	MMS	M	M	S
3	MMM	M	M	M

Kessens, at 195.

“In the third step, the language model is altered. To calculate the baseline language model the orthographic representation of the words in the training corpus is used. Because there is only one variant per word this suffices. However, when a multiple pronunciation lexicon is used during recognition and the language model is trained on the orthographic representation of the words, all variants of the same word will have equal a priori probabilities (this probability is determined by

’993 Patent		
		the language model). A drawback of this is that a sporadically occurring variant may have a high a priori probability because it is a variant of a frequently occurring word, whereas the variant should have a lower a priori probability on the basis of its occurrence. Consequently, the variant may be easily confused with other words in the lexicon. A way of reducing this confusability is to base the calculation of the language model on the phone transcription of the words instead of on the orthographic transcription, i.e. on the basis of the phone transcriptions of the corpus obtained through forced recognition. A recognition test is performed using this language model, the multiple pronunciation lexicon and the updated phone models (test condition: MMM).” Kessens, at 197. “The general procedure, described above, was employed to model within-word pronunciation variation. Pronunciation variants were automatically generated by applying a set of optional phonological rules for Dutch to the transcriptions in the baseline lexicon. The rules were applied to all words in the lexicon wherever it was possible and in no specific order, using a script in which the rules and conditions were specified. All of the variants generated by the script were added to the baseline lexicon, thus creating a multiple pronunciation lexicon. We modeled within-word variation using five optional phonological rules concerning: /n/-deletion, /r/-deletion, /t/-deletion, /@/-deletion and /@/-insertion (SAMPA 2-notation is used throughout this article). These rules were chosen according to the following four criteria.” Kessens, at 197. Wester discloses, expressly or inherently, the step of “incorporating, into the language model, pronunciation probabilities associated with respective unique labels for each different pronunciation of a word, ***wherein the respective unique label for a most frequent word indicates a special status in the language model***:” In Wester, the most frequently occurring words and variants are selected, and given a special status. *See, e.g.*, “An oft mentioned advantage of articulatory-acoustic (phonological) features in speech recognition is that these features are better suited for pronunciation modeling than a purely phone-based approach. Few studies, however, have investigated whether this claim is justified or not. In a recent

’993 Patent		
		study (Lee and Wellekens 2001) an approach to modeling pronunciation variation was described in which articulatory-acoustic features are used. Lee’s approach consists of generating a multiple variant static lexicon during training, which is dynamically adjusted during recognition. The information used to generate pronunciation variants is obtained by extracting features from the speech signal (using an approach similar to King and Taylor (2000)). The features are mapped to phones which are then connected to each other to build a pronunciation network. All possible pronunciations are generated from the network and the output is smoothed by a two-pass forced recognition. The remaining variants are stored in the static lexicon. During recognition this static lexicon is adjusted per utterance. Articulatory-acoustic features are extracted from the test material, mapped to phones, and used to select those entries from the static lexicon that best match the phonetic characteristics of a given speech signal. The selected entries constitute the dynamic lexicon, which is used for recognition. A 16% relative reduction in WER was found on TIMIT (Lamel et al. 1986) compared to their baseline system.” Wester, at 17. In addition, to the extent that Wester does not incorporate Wester2 by reference, as previously discussed with respect to limitation 17.a, which is hereby incorporated by reference, it would be obvious to a person having ordinary skill in the art to combine Wester with Wester2. Wester2 also discloses, expressly or inherently, the step of “incorporating, into the language model, pronunciation probabilities associated with respective unique labels for each different pronunciation of a word, ***wherein the respective unique label for a most frequent word indicates a special status in the language model***:” In Wester2, the most frequently occurring words and variants are selected, and given a special status. *See, e.g.*, “Many studies (e.g. Cohen (1989, Yang and Martens (2000, Ma et al. (1998)) have found that probabilities of the variants (or probabilities of rules) play an important role in whether an approach to modeling pronunciation variation is successful or not. In this study, this was once again shown by comparing results between Phicos and the ICSI system in §6.2. Not including priors in the ICSI system and not incorporating variants in the language model for Phicos showed significant deteriorations, whereas including probabilities showed significant improvements over

’993 Patent		
		the baseline. Yet if we are to relate this to the findings of McAllaster et al. (1998) and Saraclar et al. (2000): if one can accurately predict word pronunciations in a certain test utterance the performance should improve substantially, we must conclude that estimating the priors for a whole lexicon is not optimal. The point is that a good estimation of priors is probably a conditional probability with speaker, speaking mode, speaking rate, subject, etc. as conditionals. Some of these factors can be dealt with in a two-pass scheme by rescoring n-best lists as the pronunciation models in Fosler-Lussier (1999) showed; however, the gains found in this study remain small as it is extremely difficult to accurately estimate the conditionals.” Wester2, at 119. In addition, to the extent that Wester does not incorporate Kessens by reference, as previously discussed with respect to limitation 17.a, which is hereby incorporated by reference, it would be obvious to a person having ordinary skill in the art to combine Wester with Kessens. Kessens also discloses, expressly or inherently, the step of “incorporating, into the language model, pronunciation probabilities associated with respective unique labels for each different pronunciation of a word, ***wherein the respective unique label for a most frequent word indicates a special status in the language model***:” In Kessens, the 50 most frequently occurring word sequences are selected, and given a special status. *See, e.g.*, “The first step in cross-word method 1 consisted of selecting the 50 most frequently occurring word sequences from our training material. Next, from those 50 word sequences we chose those words which are sensitive to the cross-word processes cliticization, contraction and reduction. This led to the selection of seven words which made up 9% of all the words in the training corpus (see Table 2). The variants of these words were added to the lexicon and the rest of the steps of the general procedure were carried out (see Section 2.2). Table 2 shows the selected words (column 1), the total number of times the word occurs in the training material (column 2), their baseline transcriptions (column 3) and their added cross-word variants (column 4).” Kessens, at 199.

’993 Patent		
		“For all methods, the best results are obtained when pronunciation variants are used during training and recognition, and when they are added to the language model (MMM). All methods lead to an improvement in the CSR’s performance when their results are compared to the result of the baseline (SSS). These improvements are summed up in Table 5. Modeling within-word variation in isolation gives a significant improvement of 0.68%, and in combination with cross-word method 2, the improvement is also significant.” Kessens, at 204.
17.c	after incorporating the pronunciation probabilities into the language model, recognizing an utterance using the language model.	Wester discloses, expressly or inherently, the step of “after incorporating the pronunciation probabilities into the language model, recognizing an utterance using the language model:” Wester discloses testing the recognizer with these techniques as well as the resulting performance disclosed. *See, e.g.*, “Using the methods for modeling cross-word pronunciation variation, a total improvement of 0.16% was found for cross-word method 1, and 0.30% for cross-word method 2. A combination of modeling within-word and cross-word pronunciation variation leads to a total improvement of 0.61% for method 1, and a total improvement of 1.12% for crossword method 2. However, a great deal of the improvement for cross-word method 2 is due to adding multi-words (0.34%). We also investigated whether the sum of the improvements for the cross-word methods tested in isolation is comparable to the improvement obtained when testing combinations of the methods, and found that this is not the case. For crossword method 1, the sum of the methods in isolation gives better results than using the methods in combination, whereas for cross-word method 2, the combination leads to larger improvements than the sum of the results in isolation.” Wester, at 20-21. “To further understand the results that were found, we carried out a partial error analysis in which the utterances recognized with the baseline system were compared to those recognized with the experimental condition in which pronunciation variation was incorporated at all levels for a combination of within-word variants and cross-word variants modeled by multiwords. This error analysis showed that 14.7% of the recognized utterances changed, whereas a net improvement of

’993 Patent		
		only 1.3% in the sentence error rate was found (and 1.12% in the WER). Thus, the WER only reflects the net result obtained, and our error analysis showed that this is only a fraction of what actually happens due to applying our methods.” Wester, at 21. “To summarize, we obtained the best results when within-word pronunciation variation and cross-word pronunciation variation using multi-words were modeled in combination, and all the steps of the general procedure had been carried out. Using only five phonological rules and 22 multi-words, a relative improvement of 8.8% was found (12.75% -11.63%).” Wester, at 21. “The main goal of the research presented in this thesis was to improve the performance of Dutch ASR. Statistically significant improvements in WER were found, both for the knowledge-based and data-derived approaches (Kessens et al. 1999a; Wester 2001). The results presented in publication 1 and 3 show that in order to obtain significant improvements in WERs, prior probabilities for the variants should be incorporated in the recognition process in addition to adding variants to the lexicon.” Wester, at 32. “In publication 1, another of our objectives was formulated as follows: ‘Our long-term goal is to find the set of rules which is optimal for modeling pronunciation variation.’ It is difficult to conclude whether this goal has been reached or not. It is possible that in the course of the research carried out for this thesis the optimal set of variants for the VIOS data was found. However, if that is the case, it went unnoticed, as we implicitly assumed that performing recognition with the optimal set of variants would lead to lower WERs. In Section 1.7.1, I argued that the reason for the lack of improvement in WER is because conditional probabilities are not taken into account in a static lexicon. Therefore, it could be the case that we have the correct set of variants to describe the

’993 Patent		
		pronunciation variation present in the VIOS material, but that this is not reflected in the WERs because of lexical confusability.” Wester, at 32. “In Section 1.7, lexical confusability, phone transcriptions, and the beads-on-a-string paradigm were presented as shortcomings of the segmental approach to modeling pronunciation variation. This may give the impression that there is no future for pronunciation modeling. However, the outlook for pronunciation modeling is not quite that bleak. It is my impression that the future of pronunciation modeling should lie in employing different levels of linguistic information to predict and model the variation present in the speech material. This section gives a few examples of how this can be achieved in pronunciation modeling.” Wester at 32-33. “The objective of retraining the acoustic models on the basis of the output of forced alignment is not only to obtain more accurate acoustic models but also to achieve a better match between the multiple pronunciation lexicon and the acoustic models used during recognition. In various studies improvements in recognition results were found after retraining the acoustic models (Sloboda and Waibel 1996; Riley et al. 1999). However, in some studies no difference in performance was measured (Holter and Svendsen 1999), or even a deterioration was found (Beulen et al. 1998). Strik and Cucchiarini (1999) mention that these retranscriptionretraining steps can be iterated, and Saraclar (2000) and Kessens et al. (1999b) demonstrate that most of the gain is found as a result of the first iteration.” Wester, at 16. “In order to achieve this objective, we proposed a general procedure for modeling pronunciation variation. This procedure affects all three levels of the CSR at which modeling can take place: i.e. the lexicon, the phone models and the language model (Strik and Cucchiarini 1999). This means that variants were added to the lexicon and language models, and that the phone models were

’993 Patent		
		retrained on a retranscription of the training material obtained through forced alignment. The general procedure was employed to model within-word variation as well as cross-word variation.” Wester, at 20. “The main results that we found are the following. The baseline system WER is 12.75%. For the within-word method, adding pronunciation variants to the lexicon leads to an improvement of 0.31% compared to the baseline. When, in addition, retrained phone models are used, a further improvement of 0.22% is found, and finally, incorporating variants into the language model leads to a further improvement of 0.15%. In total, a small but statistically significant improvement of 0.68% was found for modeling within-word pronunciation variation.” Wester, at 20. “The results presented in publication 3 of this thesis show that simply adding a great deal of variants to the lexicon leads to a deterioration in WER. Therefore, prior probabilities are included in the decoding process. In Section 1.7.1, it was argued that although prior probabilities are important to include in the recognition process they do not suffice for modeling pronunciation variation and that conditional probabilities are possibly the key to reducing WERs.” Wester, at 33. In addition, to the extent that Wester does not incorporate Kessens by reference, as previously discussed with respect to limitation 17.a, which is hereby incorporated by reference, it would be obvious to a person having ordinary skill in the art to combine Wester with Kessens. Kessens also discloses, expressly or inherently, the step of “after incorporating the pronunciation probabilities into the language model, recognizing an utterance using the language model:” Kessens discloses testing the recognizer with these techniques as well as the resulting performance disclosed. *See, e.g.*,

'993 Patent		
		Table 4 WER for the within-word method (within), cross-word method 1 (cross 1), cross-word method 2 (cross 2), the within-word method with multi-words added to the lexicon and language model (within + multi), and the combination of the within-word method with cross-word method 1 (within + cross 1) and cross-word method 2 (within + cross 2) (table below) Kessens, at 201. "As was explained in Section 2.5, two processes play a role when using multi-words to model cross crossword pronunciation variation, i.e., firstly, adding the multi-words and, secondly, adding variants of the multi-words. To measure the effect of only adding the multi-words (without variants), the experiments for within-word variation were repeated with the multi-words added to the lexicon and the language model. Row 5 in Table 4 (within + multi) shows the results of these experiments. The effect of the multi-words can be seen by comparing these results to the results of the withinword method (row 2 in Table 4). The comparison clearly shows that adding multi-words to the lexicon and the language model leads to improvements for all conditions. The improvements range from 0.34% to 0.41% for the different conditions." Kessens, at 202. "In row 6 (within + cross 1) and row 7 (within + cross 2) of Table 4, the results of combining the within-word method with the two cross-word methods are shown. It can be seen that adding variants to the lexicon improves the CSR's performance by 0.05% and 0.04% for cross-word methods 1 and 2, respectively (SSS ^ MSS, SSS* ^ MSS). Using retrained phone models (MSS ^ MMM) improves the WER by another 0.12% for cross-word method 1, and 0.07% for cross-word method 2. Finally, the improvements are largest when the pronunciation variants are used in the

	SSS	MSS	MMS	MMM
within	12.75	12.44	12.22	12.07
cross 1	12.75	13.00	12.89	12.59
cross 2	12.41*	12.74	12.99	12.45
within + multi	12.41*	12.05	11.81	11.72
within + cross 1	12.75	12.70	12.58	12.14
within + cross 2	12.41*	12.37	12.30	11.63

* Multi-words added to the lexicon and the language model.

’993 Patent		
		language model too (MMM). For cross-word method 1, a further improvement of 0.44% is found compared to MMS, and for cross-word method 2, an even larger improvement of 0.67% is found.” Kessens, at 202. “For the combination of the within-word method with cross-word method 1, a total improvement of 0.61% is found for the test condition MMM compared to the baseline (SSS). For the same test condition, the combination of the within-word method with cross-word method 2 leads to a total improvement of 0.78% compared to the SSS* condition.” Kessens, at 202. “For all methods, the best results are obtained when pronunciation variants are used during training and recognition, and when they are added to the language model (MMM). All methods lead to an improvement in the CSR’s performance when their results are compared to the result of the baseline (SSS). These improvements are summed up in Table 5. Modeling within-word variation in isolation gives a significant improvement of 0.68%, and in combination with cross-word method 2, the improvement is also significant.” Kessens, at 204. “Up until now we have only presented our results in terms of WER (as is done in most studies). WERs give an indication of the net change in the performance of one CSR compared to another one. However, they do not provide more detailed information on how the recognition results of the two CSRs differ. Since this kind of detailed information is needed to gain more insight, we carried out a partial error analysis. To this end, we compared the utterances recognized with the baseline test to those recognized with our best test (MMM for within + cross 2 in Table 4). For the moment, we have restricted our error analysis to the level of the whole utterance, mainly for practical reasons. In the near future, we plan to do it at the word level too.” Kessens, at 204.

’993 Patent		
		“In this research, we attempted to model two types of variation: within-word variation and cross-word variation. To this end, we used a general procedure in which pronunciation variation was modeled at the three different levels in the CSR: the lexicon, the phone models and the language model. We found that the best results were obtained when all of the steps of the general procedure were carried out, i.e. when pronunciation variants were incorporated at all three levels. Below, the results of incorporating pronunciation variants at all three levels are successively discussed.” Kessens, at 205. “In the third step, pronunciation variants were also incorporated at the level of the *language model* (MMS ^ MMM), which is beneficial to all methods. Moreover, the effect of adding variants to the language model is much larger for the crossword methods than for the within-word method. This is probably due to the fact that many recognition errors introduced in the first step (see above) are corrected when variants are also included in the language model. When cross-word variants are added to the lexicon (step 1), short sequences of only one or two phones long (like e.g. the phone /k/) can easily be inserted, as was argued above. The output of forced recognition reveals that the cross-word variants occur less frequently than the canonical pronunciations present in the baseline lexicon: on average in about 13% of the cases for cross-word method 1, and 9% for cross-word method 2. In the language model with cross-word variants included, the probability of these cross-word variants is thus lower than in the original language model and, consequently, it is most likely that they will be inserted less often.” Kessens, at 205. “To summarize, we obtained the best results when within-word pronunciation variation and cross-word pronunciation variation using multiword were modeled in combination, and all the steps of the general procedure had been carried out. Using only five phonological rules and 22 multi-words a relative improvement of 8.8% was found (12.75%-11.63%).” Kessens, at 207.

’993 Patent		
Claim 19		
19	The computer-readable storage device of claim 17, wherein the language model is generated by modeling pronunciation dependencies across word boundaries.	As discussed above with respect to claim 17, Wester, either by itself, or in combination with Steinbiss, discloses, expressly or inherently, the computer-readable storage device of claim 17. *See supra* claim [17], which is incorporated by reference herein. Wester discloses, expressly or inherently, a “language model [that] is generated by modeling pronunciation dependencies across word boundaries.” Wester describes using “cross-word processes” as part of the pronunciation variations it accounts for in the language model. *See, e.g.*, “Incorporating pronunciation variation in the language model can be carried out by estimating the probabilities of the variants instead of the probabilities of the words. This is of course only possible if the pronunciation variants are transcribed in the training material, and the language models are trained on this material. An intermediate level of modeling pronunciation variation in the language model is possible in the form of word classes. In particular, this approach is taken to deal with processes of cross-word variation such as liaisons in French (Brieussel-Pousse and Perennou 1999).” Wester, at 16. “In order to achieve this objective, we proposed a general procedure for modeling pronunciation variation. This procedure affects all three levels of the CSR at which modeling can take place: i.e. the lexicon, the phone models and the language model (Strik and Cucchiarini 1999). This means that variants were added to the lexicon and language models, and that the phone models were retrained on a retranscription of the training material obtained through forced alignment. The general procedure was employed to model within-word variation as well as cross-word variation.” Wester, at 20. “A limited number of cross-word processes were modeled, using two different techniques. The type of cross-word processes we focussed on were cliticization, reduction and contraction (Booij 1995). The first technique consisted of modeling cross-word processes by adding the cross-word

’993 Patent		
		variants directly to the lexicon (cross-word method 1), and in the second approach this was done by using multi-words (cross-word method 2). These cross-word approaches were each tested in isolation and in combination with the set of within-word variants (all five rules).” Wester, at 20. “Using the methods for modeling cross-word pronunciation variation, a total improvement of 0.16% was found for cross-word method 1, and 0.30% for cross-word method 2. A combination of modeling within-word and cross-word pronunciation variation leads to a total improvement of 0.61% for method 1, and a total improvement of 1.12% for crossword method 2. However, a great deal of the improvement for cross-word method 2 is due to adding multi-words (0.34%).” Wester, at 20. “The main goal of the research presented in this thesis was to improve the performance of Dutch ASR. Statistically significant improvements in WER were found, both for the knowledge-based and data-derived approaches (Kessens et al. 1999a; Wester 2001). The results presented in publication 1 and 3 show that in order to obtain significant improvements in WERs, prior probabilities for the variants should be incorporated in the recognition process in addition to adding variants to the lexicon.” Wester, at 32. “In publication 1, another of our objectives was formulated as follows: ‘Our long-term goal is to find the set of rules which is optimal for modeling pronunciation variation.’ It is difficult to conclude whether this goal has been reached or not. It is possible that in the course of the research carried out for this thesis the optimal set of variants for the VIOS data was found. However, if that is the case, it went unnoticed, as we implicitly assumed that performing recognition with the optimal set of variants would lead to lower WERs. In Section 1.7.1, I argued that the reason for the lack of improvement in WER is because conditional probabilities are not taken into account in a static lexicon. Therefore, it could be the case that we have the correct set of variants to describe the

'993 Patent		
		pronunciation variation present in the VIOS material, but that this is not reflected in the WERs because of lexical confusability." Wester, at 32. In addition, to the extent that Wester does not incorporate Kessens by reference, as previously discussed with respect to limitation 17.a, which is hereby incorporated by reference, it would be obvious to a person having ordinary skill in the art to combine Wester with Kessens. Kessens also discloses, expressly or inherently, a "language model [that] is generated by modeling pronunciation dependencies across word boundaries." Kessens describes using "cross-word variants," which Wester integrates into the language model. *See, e.g.*, "This article describes how the performance of a Dutch continuous speech recognizer was improved by modeling pronunciation variation. We propose a general procedure for modeling pronunciation variation. In short, it consists of adding pronunciation variants to the lexicon, retraining phone models and using language models to which the pronunciation variants have been added. First, within-word pronunciation variants were generated by applying a set of five optional phonological rules to the words in the baseline lexicon. Next, a limited number of cross-word processes were modeled, using two different methods. In the first approach, cross-word processes were modeled by directly adding the cross-word variants to the lexicon, and in the second approach this was done by using multi-words. Finally, the combination of the within-word method with the two cross-word methods was tested. The word error rate (WER) measured for the baseline system was 12.75%. Compared to the baseline, a small but statistically significant improvement of 0.68% in WER was measured for the within-word method, whereas both cross-word methods in isolation led to small, non-significant improvements. The combination of the within-word method and cross-word method 2 led to the best result: an absolute improvement of 1.12% in WER was found compared to the baseline, which is a relative improvement of 8.8% in WER." Kessens, at Abstract.

’993 Patent		
		“In this research, we attempted to model two types of variation: within-word variation and cross-word variation. To this end, we used a general procedure in which pronunciation variation was modeled at the three different levels in the CSR: the lexicon, the phone models and the language model. We found that the best results were obtained when all of the steps of the general procedure were carried out, i.e. when pronunciation variants were incorporated at all three levels. Below, the results of incorporating pronunciation variants at all three levels are successively discussed.” Kessens, at 205. “In the first step, variants were only incorporated at the level of the *lexicon.* Compared to the baseline (SSS ^ MSS), an improvement was found for the within-word method and for the within-word method in combination with each of the two crossword methods. However, a deterioration was found for the two cross-word methods in isolation. A possible explanation for the deterioration for cross-word method 1 is related to the fact that the pronunciation variants of cross-word method 1 are very short (see Table 2); some of them consist of only one phone. Such short variants can easily be inserted; for instance, the plosives /k/ and /t/ might occasionally be inserted at places where clicks in the signal occur. Furthermore, this effect is facilitated by the high frequency of occurrence of the words involved, i.e. they are favored by the language model. Similar things might happen for cross-word method 2. Let us give an example to illustrate this: A possible variant of the multi-word “ik_wil” /IkwIl/ is /kwIl/. The latter might occasionally be confused with the word “wil” /wIl/. This confusion leads to a substitution, but effectively it is the insertion of the phone /k/. Consequently, insertion of /k/ and other phones is also possible in cross-word method 2, and this could explain the deterioration found for cross-word method 2.” Kessens, at 205. “When, in the second step, pronunciation variation is also incorporated at the level of the *phone models* (MSS ^ MMS), the CSR’s performance improved in all cases, except in the case of crossword method 2. A possible cause of this deterioration in performance could be that the phone models were not retrained properly. During forced recognition, the option for recognizing a pause between the separate parts of the multi-words was not given. As a consequence, if a pause occurred in the acoustic signal of a multi-word, the pause was used to train the surrounding phone

’993 Patent		
		models, which results in contaminated phone models. Error analysis revealed that in 5% of the cases a pause was indeed present within the multi-words in our training material. Further research will have to show whether this was the only cause of the deterioration in performance or whether there are other reasons why retraining phone models using multi-words did not lead to improvements.” Kessens, at 205. “In the third step, pronunciation variants were also incorporated at the level of the *language model* (MMS ^ MMM), which is beneficial to all methods. Moreover, the effect of adding variants to the language model is much larger for the crossword methods than for the within-word method. This is probably due to the fact that many recognition errors introduced in the first step (see above) are corrected when variants are also included in the language model. When cross-word variants are added to the lexicon (step 1), short sequences of only one or two phones long (like e.g. the phone /k/) can easily be inserted, as was argued above. The output of forced recognition reveals that the cross-word variants occur less frequently than the canonical pronunciations present in the baseline lexicon: on average in about 13% of the cases for cross-word method 1, and 9% for cross-word method 2. In the language model with cross-word variants included, the probability of these cross-word variants is thus lower than in the original language model and, consequently, it is most likely that they will be inserted less often.” Kessens, at 205. “One of the questions we posed in the introduction was what the best way of modeling crossword variation is. On the basis of our results we can conclude that when cross-word variation is modeled in isolation, cross-word method 2 performs better than cross-word method 1, but the difference is non-significant. In combination with the within-word method, cross-word method 2 leads to an improvement compared to the within-word method in isolation. This is not the case for cross-word method 1, which leads to a degradation in WER. Therefore, it seems that cross-word method 2 is more suitable for modeling cross-word pronunciation variation. It should be noted, however, that most of the improvements gained with cross-word method 2 are due to adding the multi-words to the lexicon and the language model. An explanation for these improvements is that by adding

’993 Patent		
		multi-words to the language model the span of the unigram and bigram increases for the most frequent word sequences in the training corpus. Thus, more context information can be used during the recognition process. Furthermore, it should also be noted that only a small amount of data was involved in the cross-word processes which were studied; only 6-9% of the words in the training corpus were affected by these processes. Therefore, we plan to test cross-word methods 1 and 2 for a larger amount of data and a larger number of cross-word processes.” Kessens, at 205-06. “To summarize, we obtained the best results when within-word pronunciation variation and cross-word pronunciation variation using multiword were modeled in combination, and all the steps of the general procedure had been carried out. Using only five phonological rules and 22 multi-words a relative improvement of 8.8% was found (12.75%-11.63%).” Kessens, at 207.

APPENDIX B-2
Invalidity Claim Chart for U.S. Patent No. 8,532,993 ('993 patent)
Florian Schiel, et al., *Statistical Modelling of Pronunciation: It's not the Model, It's the Data*, 1998 ("Schiel") [1]

On October 16, 2020, Nuance narrowed the asserted '993 patent claims to 17 and 19. Nuance's Disclosure Limiting Asserted Claims of Phase 1 Patents. Claims 17 and 19 of the '993 patent are anticipated and/or rendered obvious by Schiel alone or in combination with at least the following references. The chart below discloses how prior art references identified by Omilia disclose, either expressly or inherently, and/or render obvious each of the asserted claims:

(1) Steinbiss et al., *The Philips research system for large-vocabulary continuous-speech recognition*, Proc. of 3rd European Conference on Speech Communication and Technology EUROSPEECH '93, 2125-2128 (1993) ("Steinbiss")[2]

(2) Jain, et al., *Creating Speaker-Specific Phonetic Templates with a Speaker-Independent Phonetic Recognizer: Implications for Voice Dialing*, IEEE International Conference on Acoustics, Speech, and Signal Processing Conference Proceedings, 881-884 (1996) ("Jain")[3]

(3) Kessens et al., *Improving the performance of a Dutch CSR by modeling within-word and cross-word pronunciation variation*, Speech Communication 29 (1999) 193-207 ("Kessens")[4]

This invalidity claim chart is based on Omilia's present understanding of the asserted claims, the Court's August 6, 2020 Claim Construction Order (ECF. No. 241), Plaintiff ("Nuance's") initial and all subsequent supplemental Infringement Contentions, its Response to Omilia NLS' Supplemental Preliminary Non-Infringement and Invalidity Contentions, its Response to Omilia NLS' Interrogatory No. 9, and Omilia's investigation to date. Omilia is not adopting Nuance's apparent constructions, nor is Omilia admitting to the accuracy of any particular construction. Omilia reserves all rights to amend this invalidity claim chart if Nuance further amends its Infringement Contentions, or in response to any statements made by Nuance at any phase of this litigation. In addition, Discovery is still ongoing, and Omilia reserves its right to supplement its contentions with additional evidence as discovery continues.

The citations provided are exemplary and do not necessarily include each and every disclosure of the limitation in the references. Omilia has cited the most relevant portions of the identified prior art. Other portions of the identified prior art may additionally disclose, either

---

[1] Schiel was published in May 1998. Schiel is prior art under at least pre-AIA 35 U.S.C. § 102(a), and (b).
[2] Steinbiss was published in September, 1993. Steinbiss is prior art under at least pre-AIA 35 U.S.C. § 102(a), and (b).
[3] Jain was published in 1996. Jain is prior art under at least pre-AIA 35 U.S.C. § 102(a), and (b).
[4] Kessens was published in 1999. Kessens is prior art under at least pre-AIA 35 U.S.C. § 102(a), and (b).

expressly or inherently, and/or render obvious one or more limitations of the asserted claims. To provide context or aid in understanding the prior art and the state of the art at the time of the alleged invention, Omilia reserves the right to rely on: (1) uncited portions of the identified prior art; (2) other prior art not identified herein; (3) references that show the state of the art (irrespective of whether such references themselves qualify as prior art to the asserted patents); (4) factual testimony from the inventors or authors of the prior art references, or purveyors of prior art devices; and/or (5) expert testimony.

Citations to a particular drawing or figure in the accompanying charts should be understood to include the description of the drawing or figure, as well as any text associated with the drawing or figure. Citations to particular text concerning a drawing or figure, should also be understood to encompass that drawing or figure as well. The identified prior art expressly and/or inherently discloses the features of the asserted claims under all proposed constructions of the asserted claims. To establish the inherency of certain features of the prior art to invalidate the asserted claims, Omilia may rely on cited or uncited portions of the prior art, other documents, factual testimony, and expert testimony. To the extent that the Court finds that this reference does not explicitly teach certain limitations in the asserted claims, such limitations would have been inherent and/or obvious and Omilia reserves the right to supplement or amend this claim chart to address such a finding.

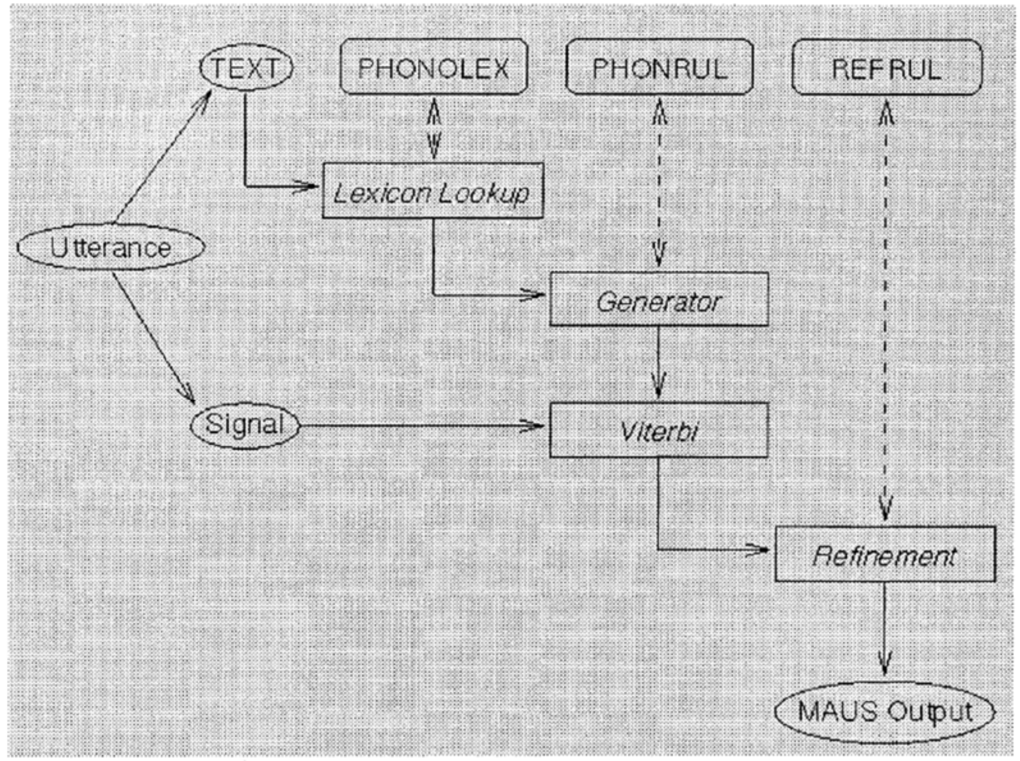
’993 Patent		
Claim 17		
17.pre	A computer-readable storage device having instructions stored which, when executed on a processor, cause the processor to perform operations comprising:	To the extent that the preamble is limiting, Schiel discloses, expressly or inherently, “computer-readable storage device having instructions stored which, when executed on a processor, cause the processor to perform operations:” In Schiel, the process described was developed and implemented on the MAUS computer system using conventional computer elements. *See, e.g.*, “The MAUS system was developed at the Bavarian Archive for Speech Signals (BAS) to facilitate the otherwise very time-consuming manual labelling and segmentation of speech corpora into phonetic units. Initially funded by the German government within the VERBMOBIL I project, MAUS is now further extended by BAS with the aim to automatically improve all BAS speech corpora by means of complete broad phonetic transcriptions and segmentations. The basic motivation for MAUS is the hypothesis that automatic speech recognition (ASR) of conversational speech as well as high quality 'concept-to-speech' systems

’993 Patent		
		will require huge amounts of carefully labelled and segmented speech data for their successful progress.” Schiel, at 131. “Input to the MAUS system is the digitized speech wave and any kind of orthographic representation that reflects the chain of words in the utterance. Optionally there might be markers for non-speech events as well, but this is not essential for MAUS. The output of MAUS is a sequence of phonetic/phonemic symbols from the extended German SAM Phonetic Alphabet ([3]) together with the time position within the corresponding speech signal.” Schiel, at 131. In addition, it would be obvious to a person having ordinary skill in the art to combine Schiel with Steinbiss. Both Schiel and Steinbiss are in the same field of art. A person having ordinary skill in the art would be motivated to combine Schiel with Steinbiss. A person having ordinary skill in the art considering Schiel would be motivated to implement the system on a computer in order to construct a working system, as disclosed by Steinbiss. A person having ordinary skill in the art would have a reasonable expectation of success in implementing Schiel on the computer of Steinbiss because the combination involves the predictable use of prior art elements according to their established functions. The motivation to combine these references would at least include: • Combining prior art elements according to known methods to yield predictable results (using a computer to perform speech recognition using computers was well known in the art); • Simple substitution of one known element for another to obtain predictable results (implementing the system of Schiel on a computer, as disclosed by Steinbiss);

’993 Patent		
		• Use of known technique to improve similar devices (methods, or products) in the same way (the technique of using a computer to perform speech recognition was known); • Applying a known technique to a known device (method, or product) ready for improvement to yield predictable results (using a computer to perform speech recognition using computers was well known in the art); • “Obvious to try”—choosing from a finite number of identified, predictable solutions, with a reasonable expectation of success (Schiel teaches that the system uses digitized speech, which can be processed using a computer); • Market forces and benefits associated with the known benefits of automatic speech recognition. Teaching of prior art would have lead a POSA to combine the references to arrive at a speech recognition implemented on a generic computer. Steinbiss also discloses, expressly or inherently, “computer-readable storage device having instructions stored which, when executed on a processor, cause the processor to perform operations: In Steinbiss, the process described was developed and implemented using PC based implementation. *See, e.g.*, “The system has been successfully applied to the American English DARPA RM task. Here, we report experimental results for a German 13 000-word Philips internal dictation task. In addition to the scientific prototype, a PC version has been set up which is described here for the first time.” Steinbiss, at 2125. “The organization of the paper is as follows. We first summarize the statistical approach to speech recognition and then describe the four main entities of our system: acoustic analysis, acoustic phonetic modelling, language modelling and search. A section with experiments on our

’993 Patent		
		internal dictation task follows. The final section describes a PC based implementation of our system.” Steinbiss, at 2125. “8. A PC Based Continuous Speech-Recognition System for Dictation . . . The system developed at Philips Dictation Systems, Vienna, and described here adopts this non-interactive approach and thus allows the person to dictate with a natural speaking style. After the speech is processed by the speech recognizer, the secretary has only to correct the recognition errors, which is both faster and a more interesting job to do.” Steinbiss, at 2128. “Speech recognition runs remotely on a PC which is connected to the network. An acoustic front-end performs the acoustic analysis. Recognition is sped up by a dedicated co-processor board containing application-specific ICs. Depending on the speaker and the specific boundary conditions, recognition with a 10 - 20 000-word vocabulary runs in 1 - 3 times real-time.” Steinbiss, at 2128.
17.a	approximating transcribed speech using a phonemic transcription dataset associated with a speaker, to yield a language model, where the phonemic transcription dataset is based on a pronunciation model of the speaker;	Schiel discloses, expressly or inherently, the step of “approximating transcribed speech using a phonemic transcription dataset associated with a speaker, to yield a language model, where the phonemic transcription dataset is based on a pronunciation model of the speaker:” Schiel discloses, expressly or inherently, the step of “***approximating transcribed speech using a phonemic transcription dataset associated with a speaker***, to yield a language model, ***where the phonemic transcription dataset is based on a pronunciation model of the speaker***.” In Schiel, a recognition system is trained using digitized speech waves that are automatically segmented and converted into phonetic/phonemic symbols. A POSA would further understand that the digitized speech wave is associated with a particular speaker. A POSA would further

’993 Patent		
		understand that the process of transforming the “digitized speech wave” to “a sequence of phonetic/phonemic symbols” disclosed by Schiel is “based on a pronunciation model of the speaker.” *See e.g.*, “Input to the MAUS system is the digitized speech wave and any kind of orthographic representation that reflects the chain of words in the utterance. Optionally there might be markers for non-speech events as well, but this is not essential for MAUS. The output of MAUS is a sequence of phonetic/phonemic symbols from the extended German SAM Phonetic Alphabet ([3]) together with the time position within the corresponding speech signal.” Schiel, at 131.

’993 Patent		
		 **Figure 1. The MAUS system - block diagram** Schiel, at Fig. 1. “In a first step the orthographic string of the utterance is looked up in a canonical pronunciation dictionary (e.g. PHONOLEX, see [14]) and processed into a Markov chain (represented as a directed acyclic graph) containing all possible alternative pronunciations using either a set of data driven microrules or using the phonetic expert system PHONRUL.” Schiel, at 132.

’993 Patent		
		“The second stage of MAUS is a standard HMM Viterbi alignment where the search space is constrained by the directed acyclic graph from the first stage (see figure 2 for an example). Currently we use the HTK 2.0 as the aligner ([12]) with the following preprocessing: 12 MFCCs + log Energy, Delta, Delta-delta every 10 msec. Models are left-right, 3 to 5 states and 5 mixtures per state. No tying of parameters was applied to keep the model as sharp as possible. The models were trained to manually segmented speech only (no embedded re-estimation).” Schiel, at 133. “The outcome of the alignment is a transcript and a segmentation of 10 msec accuracy, which is quite broad. Therefore in a third stage REFINE the segmentation is refined by a rule-based system working on the speech wave as well as on other fine-grained features. However, the third stage cannot alter the transcript itself, only the individual segment boundaries.” Schiel, at 133. “In terms of accuracy of segment boundaries the comparison between manual segmentations shows a high agreement: on average 93% of all corresponding segment boundaries deviate less than 20msec from each other. The average percentage of corresponding segment boundaries in a MAUS versus a manual segmentation is only 84%. This yields a relative performance of 90.3%. We hope that a further improvement of the third stage of MAUS will increase these already encouraging results.” Schiel, at 133. “We trained and tested the recogniser with the same amount of data in two different fashions: • *Baseline System* Standard bootstrapping to manually labelled data (1h40) and iterative embedded re-estimation (segmental-k-means) using 30h of speech until the performance on the independent test set converged. The re-estimation process used a canonical

’993 Patent		
		pronunciation dictionary with one pronunciation per lexical entry. The system was tested with the same canonical dictionary. • *MAUS System* This system was bootstrapped to one third of the training corpus (approx. 10h of speech) using the MAUS segmentation and then iteratively re-estimated (30h of speech) using not the canonical dictionary but the transcripts of the MAUS analysis (note that the segmental information of the MAUS analysis is NOT used for the re-estimation). The system was tested with the probabilistic pronunciation model described in section 2.1. using the pruning parameters $N = 20$ and $M = 0\%$.” Schiel, at 135. In addition, it would be obvious to a person having ordinary skill in the art to combine Schiel with Jain. Both Schiel and Jain are in the same field of art. A person having ordinary skill in the art would be motivated to combine Schiel with Jain. A person having ordinary skill in the art considering Schiel’s disclosure that it is beneficial to have multiple variants captured in the training data would be motivated to seek out a system that automatically captures a phonetic representation of pronunciation variants, as disclosed by Jain. A person having ordinary skill in the art would have a reasonable expectation of success in implementing Schiel using the phonetic transcriptions provided by Jain because the combination involves the predictable use of prior art elements according to their established functions. The motivation to combine these references would at least include: • Combining prior art elements according to known methods to yield predictable results (using speaker-dependent training data to improve accuracy for a user); • Simple substitution of one known element for another to obtain predictable results (adding additional pronunciation variants to improve accuracy of a speech recognition system);

’993 Patent		
		• Use of known technique to improve similar devices (methods, or products) in the same way (using speaker-dependent pronunciation data to improve the accuracy of a speech recognition system); • Applying a known technique to a known device (method, or product) ready for improvement to yield predictable results (using speaker-dependent pronunciation was well known in the art); • “Obvious to try”—choosing from a finite number of identified, predictable solutions, with a reasonable expectation of success (there are a finite number of known techniques to improve accuracy for speech recognition systems); • Market forces and benefits associated with the known benefits of automatic speech recognition; and • Teaching of prior art would have lead a POSA to combine the references to arrive at a system that includes transcription data reflecting pronunciation data from a speaker. Jain also discloses, expressly or inherently, the step of “***approximating transcribed speech using a phonemic transcription dataset associated with a speaker***, to yield a language model, ***where the phonemic transcription dataset is based on a pronunciation model of the speaker***.” In Jain, speaker-specific models are used to transcribe telephone speech into its phonetic subparts. A POSA would further understand that the “telephone speech” is associated with a particular speaker. *See, e.g.*, “In this section, we describe the method for generating the speaker-specific models. Our speaker-specific models represent utterances as sequences of phonemes rather than acoustic parameters. A speaker-independent phonetic front end is used to generate the string of phonemes. For this approach to work, all that is needed is that the phonetic frontend behave in a consistent manner, i.e. it must produce the same (or nearly the same) string of phonemes each time the word is spoken.”

’993 Patent		
		Jain, at 881. “The following steps produce a phonetic transcription: 1. Data Capture: Telephone speech is sampled at 8 kHz. Unnecessary silence at the beginning and end of the utterance is removed by end-point detection. 2. Feature Extraction: Seventh order RASTA features [2] are computed for every 10ms of speech using a 10ms window. This yields eight coefficients per frame.” Jain, at 881. “Another method is to force-align each phoneme string (which was generated using the template generation technique) with all the other waveforms corresponding to the same label and compute the average Viterbi score. The template with the maximum average score is selected as the word-model for that label. Table 1 shows the error rates obtained with forced-alignment. It can be seen that the error rates decreased substantially for all the 4 channels. <table><tr><th>Telephone Channel</th><th>1</th><th>2</th><th>3</th><th>4</th></tr><tr><td>Baseline</td><td>7.9</td><td>12.0</td><td>8.7</td><td>18.5</td></tr><tr><td>Forced Alignment</td><td>4.8</td><td>8.0</td><td>5.8</td><td>15.5</td></tr></table> **Table 1. Effect of Forced Alignment** ” Jain, at 882-83. “It is well known that speaker dependent systems perform better than speaker-independent systems. In several commercial products, speaker adaptation is used to adapt speaker-independent models to the current user.” Jain at 883. Schiel discloses, expressly or inherently, the step of “approximating transcribed speech using a phonemic transcription dataset associated with a speaker, ***to yield a language model***, where the

’993 Patent		
		phonemic transcription dataset is based on a pronunciation model of the speaker.” In Schiel, a recognition rule set is created using the previously-segmented digitized speech. *See e.g.*, “The usage of direct statistics like in the previous section has the disadvantage that because of lack of data most of the words will be modelled by only one variant, which in many cases will be the canonical pronunciation. An easy way to generalise to less frequent words (or unseen words) is to use not the statistics of the variants itself but the underlying rules that were applied during the segmentation process of MAUS. Note that this has nothing to do with the statistical weights of the microrules mentioned earlier in this paper; it’s the number of applications of these rules that counts. Since there is formally no distinction between microrules for segmentation in MAUS and probabilistic rules for recognition, we can use the same format and formalism for this approach as in MAUS. The step-by-step procedure is as follows: • Derive a set of statistical microrules from a subset of manually segmented data or use the rule set PHONRUL as *labelling rule set* (see section 1.). • Use the *labelling rule set* to label and segment the training corpus and count all appliances of each rule forming the statistics of the *recognition rule set*. • Apply the *recognition rule set* during the ASR search to all intra-word and inter-word phoneme strings to create statistically weighted alternate paths in the search space” Schiel, at 134. “This approach has the advantage that the statistics are more compact, independent of the dictionary used for recognition (which for sure will contain words that were never seen in the training set) and generalise knowledge about pronunciation to unseen cases. However, the last point may be a source of uncertainty, since it cannot be foreseen whether the generalisation is valid to all cases where the context matches. We cannot be sure that the context we are using is sufficient to justify the usage of a certain rule in all places where this context occurs.” Schiel, at 134.

’993 Patent		
17.b	incorporating, into the language model, pronunciation probabilities associated with respective unique labels for each different pronunciation of a word, wherein the respective unique label for a most frequent word indicates a special status in the language model; and	Schiel discloses, expressly or inherently, the step of “incorporating, into the language model, pronunciation probabilities associated with respective unique labels for each different pronunciation of a word, wherein the respective unique label for a most frequent word indicates a special status in the language model.” Schiel discloses, expressly or inherently, the step of “incorporating, into the language model, ***pronunciation probabilities associated with respective unique labels for each different pronunciation of a word***, ***wherein the respective unique label for a most frequent word indicates a special status in the language model***.” In Schiel, pronunciation probabilities are associated with unique labels for different pronunciations. Those various pronunciations result from a pruning process described in Schiel, which selects the most frequent words and identifies a special status. *See, e.g.*, “Since in the MAUS output each segment is assigned to a word reference level (Partitur Format, see [13]), it is quite easy to derive all observed pronunciation variants from a corpus and collect them in a PHONOLEX ([14]) style dictionary. The analysis of the training set of the 1996 VERBMOBIL evaluation (volumes 1-5,7,12) led to a collection of approx. 230.000 observations. The following shows a random excerpt of the resulting dictionary:

’993 Patent

```
terminlich
adj
t E 6 m i: n l I C
t E 6 m i: n I C        3
t @ m i: l I C  3
t E 6 m i: n l I C      10
t E 6 m i: l I C        1
t @ m i: n l I C        7
&
...
Karfreitag
nou
k a: 6 f r aI t a: k
k a: 6 f r aI t a: k    15
k a: 6 f r aI t a x     3
&
...
weil
par
v aI l
v a l   11
v aI    108
v aI l  207
&
...
siebenundzwanzigsten
adj
z i: b @ n U n t t s v a n t s I C s t @ n
z i: b @ n U n s v a n t s I s t @ n    1
z i: b m U n s v a n t s I k s t n      2
z i: b m U n s v a n s I C s t n        1
z i: b m U n s v a n t s I C s t @ n    1
z i: m U n s v a n t s s t @ n 1
z i: m U n s v a n t s s n     1
... (remaining 48 variants deleted)
&
...
Namen
nou
n a: m @ n
n a: m  30
n a: m @ n      15
&
...
Essen
nou
Q E s @ n
@ s n   2
E s n   16
E s @ n 6
s n     3
E s     1
Q E s @ n       7
Q E s   1
Q E s n 21
&
```
”

Schiel, at 134.

“Obviously many of the observations are not frequent enough for a statistical parameterisation. Therefore we prune the baseline dictionary in the following way:

• Observations with a total count of less than *N* per lexical item are discarded.

’993 Patent		
		• From the remaining observations for each lexical word *L* the a-posteriori probabilities *P(VIL)* that the variant *V* was observed are calculated. All variants that have less than *M%* of the total probability mass are discarded. • The remaining variants are re-normalised to a total probability mass of 1.0. Applied to the above example this yields the following more compact statistics (pruning parameters: *N* = 20, *M* = 10):

```
terminlich              0.434783
t E 6 m i: n l I C
terminlich              0.130435
t E 6 m i: n I C
terminlich              0.304348
t @ m i: n l I C
terminlich              0.130435
t @ m i: l I
Karfreitag              1.000000
k a: 6 f r aI t a: k
weil                    0.342857
v aI
weil                    0.657143
v aI l
siebenundzwanzigsten    0.509091
z i: b m U n s v a n t s I s t n
siebenundzwanzigsten    0.490909
z i: m U n s v a n t s I s t n
Namen                   0.333333
n a: m @ n
Namen                   0.666667
n a: m
Essen                   0.320000
E s n
Essen                   0.420000
Q E s n
Essen                   0.120000
E s @ n
Essen                   0.140000
Q E s @ n
```

where the second column contains the a-posteriori probabilities. This form can be directly used in a standard ASR system with multi pronunciation dictionary like HTK (version 2.1).”

Schiel, at 134.

Schiel discloses, expressly or inherently, the step of “***incorporating, into the language model***, ***pronunciation probabilities*** associated with respective unique labels for each different pronunciation of a word, wherein the respective unique label for a most frequent word indicates a special status in the language model.” In Schiel, the “recognition rule set”

’993 Patent		
		incorporates “the underlying rules” which includes the pronunciation probabilities described in the lexicon. *See, e.g.*, “The usage of direct statistics like in the previous section has the disadvantage that because of lack of data most of the words will be modelled by only one variant, which in many cases will be the canonical pronunciation. An easy way to generalise to less frequent words (or unseen words) is to use not the statistics of the variants itself but the underlying rules that were applied during the segmentation process of MAUS. Note that this has nothing to do with the statistical weights of the microrules mentioned earlier in this paper; it’s the number of applications of these rules that counts. Since there is formally no distinction between microrules for segmentation in MAUS and probabilistic rules for recognition, we can use the same format and formalism for this approach as in MAUS. The step-by-step procedure is as follows: • Derive a set of statistical microrules from a subset of manually segmented data or use the rule set PHONRUL as *labelling rule set* (see section 1.). • Use the *labelling rule set* to label and segment the training corpus and count all appliances of each rule forming the statistics of the *recognition rule set*. • Apply the *recognition rule set* during the ASR search to all intra-word and inter-word phoneme strings to create statistically weighted alternate paths in the search space” Schiel, at 134. “This approach has the advantage that the statistics are more compact, independent of the dictionary used for recognition (which for sure will contain words that were never seen in the training set) and generalise knowledge about pronunciation to unseen cases. However, the last point may be a source of uncertainty, since it cannot be foreseen whether the generalisation is valid to all cases where the context matches. We cannot be sure that the context we are using is sufficient to justify the usage of a certain rule in all places where this context occurs.” Schiel, at 134.

<table>
<tr><th colspan="3"><u>’993 Patent</u></th></tr>
<tr><td>17.c</td><td>after incorporating the pronunciation probabilities into the language model, recognizing an utterance using the language model.</td><td>Schiel discloses, expressly or inherently, the step of “after incorporating the pronunciation probabilities into the language model, recognizing an utterance using the language model.” Schiel discloses testing the recognizer with these rules as well as the resulting performance disclosed. See, e.g.,

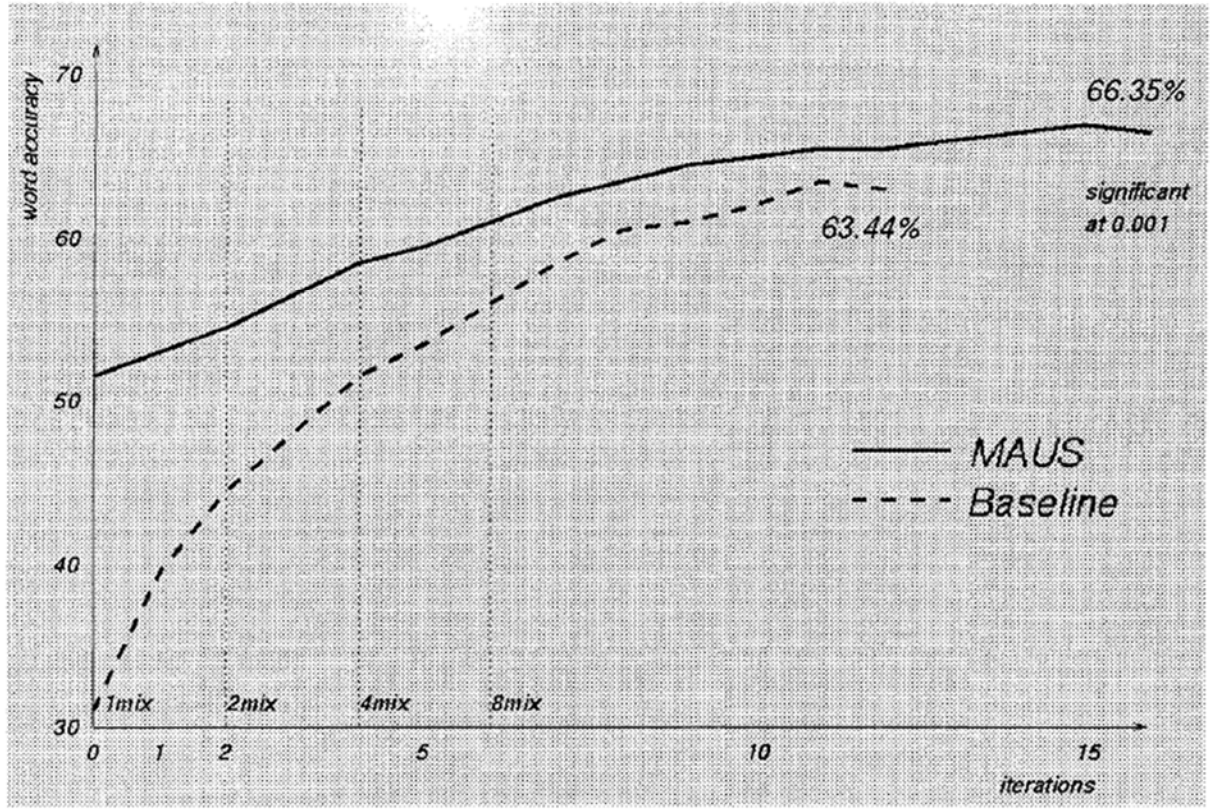
“We trained and tested the recogniser with the same amount of data in two different fashions:

• Baseline System
Standard bootstrapping to manually labelled data (1h40) and iterative embedded re-estimation (segmental-k-means) using 30h of speech until the performance on the independent test set converged. The re-estimation process used a canonical pronunciation dictionary with one pronunciation per lexical entry. The system was tested with the same canonical dictionary.
• MAUS System
This system was bootstrapped to one third of the training corpus (approx. 10h of speech) using the MAUS segmentation and then iteratively re-estimated (30h of speech) using not the canonical dictionary but the transcripts of the MAUS analysis (note that the segmental information of the MAUS analysis is NOT used for the re-estimation).
The system was tested with the probabilistic pronunciation model described in section 2.1. using the pruning parameters N = 20 and M = 0%.”

Schiel, at 135.

“Figure 3 shows the performance of both systems during the training process. Note that the MAUS system starts with a much higher performance because it was bootstrapped to 10h of MAUS data (compared to lh40min of manually labelled data for the baseline system). After training, the MAUS system converges on a significantly higher performance level of 66.35% compared to 63.44% of the baseline svstem.”

Schiel, at 135.</td></tr>
</table>

’993 Patent		
		 **Figure 3. Performance of baseline system compared to the system trained with MAUS data and probabilistic pronunciation model** Schiel, at Fig. 3.

<table>
<tr><th colspan="3"><u>’993 Patent</u></th></tr>
<tr><td colspan="3">Claim 19</td></tr>
<tr><td>19</td><td>The computer-readable storage device of claim 17, wherein the language model is generated by modeling pronunciation dependencies across word boundaries.</td><td>As discussed above with respect to claim 17, Schiel, either by itself, or in combination as described above, discloses, expressly or inherently, the computer-readable storage device of claim 17. See supra claim [17], which is incorporated by reference herein.

Schiel discloses, expressly or inherently, a “language model [that] is generated by modeling pronunciation dependencies across word boundaries.” Schiel describes using “inter-word phoneme strings” as part of its recognition rules. See, e.g.,

“The usage of direct statistics like in the previous section has the disadvantage that because of lack of data most of the words will be modelled by only one variant, which in many cases will be the canonical pronunciation. An easy way to generalise to less frequent words (or unseen words) is to use not the statistics of the variants itself but the underlying rules that were applied during the segmentation process of MAUS. Note that this has nothing to do with the statistical weights of the microrules mentioned earlier in this paper; it’s the number of applications of these rules that counts. Since there is formally no distinction between microrules for segmentation in MAUS and probabilistic rules for recognition, we can use the same format and formalism for this approach as in MAUS. The step-by-step procedure is as follows:

• Derive a set of statistical microrules from a subset of manually segmented data or use the rule set PHONRUL as labelling rule set (see section 1.).
• Use the labelling rule set to label and segment the training corpus and count all appliances of each rule forming the statistics of the recognition rule set.
• Apply the recognition rule set during the ASR search to all intra-word and inter-word phoneme strings to create statistically weighted alternate paths in the search space”

Schiel, at 134.

In addition, it would be obvious to a person having ordinary skill in the art to combine Schiel with Kessens. Both Schiel and Kessens are in the same field of art. A person having ordinary</td></tr>
</table>

’993 Patent		
		skill in the art would be motivated to combine Schiel with Kessens. A person having ordinary skill in the art considering Schiel’s disclosure recognition rules may be applied to “inter-word phoneme strings” would be motivated to seek out a system that may be used to generate such strings, as disclosed by Kessens. A person having ordinary skill in the art would have a reasonable expectation of success in implementing Schiel using the phonetic transcriptions provided by Kessens because the combination involves the predictable use of prior art elements according to their established functions. The motivation to combine these references would at least include: • Combining prior art elements according to known methods to yield predictable results (using cross-word pronunciations to yield inter-word phoneme strings); • Simple substitution of one known element for another to obtain predictable results (adding cross-word pronunciations for a recognition rule set); • Use of known technique to improve similar devices (methods, or products) in the same way (using cross-word pronunciation data to create a recognition rule set); • Applying a known technique to a known device (method, or product) ready for improvement to yield predictable results (using cross-word pronunciation data was well known in the art); • “Obvious to try”—choosing from a finite number of identified, predictable solutions, with a reasonable expectation of success (there are a finite number of known techniques to produce inter-word recognition rule sets); • Market forces and benefits associated with the known benefits of automatic speech recognition; and • Teaching of prior art would have lead a POSA to combine the references to arrive at a system that includes cross-word pronunciation dependencies.

’993 Patent		
		Kessens also discloses, expressly or inherently, a “language model [that] is generated by modeling pronunciation dependencies across word boundaries.” Kessen describes using “cross-word variation” in its pronunciation variants. *See, e.g.*, “This article describes how the performance of a Dutch continuous speech recognizer was improved by modeling pronunciation variation. We propose a general procedure for modeling pronunciation variation. In short, it consists of adding pronunciation variants to the lexicon, retraining phone models and using language models to which the pronunciation variants have been added. First, within-word pronunciation variants were generated by applying a set of five optional phonological rules to the words in the baseline lexicon. Next, a limited number of cross-word processes were modeled, using two different methods. In the first approach, cross-word processes were modeled by directly adding the cross-word variants to the lexicon, and in the second approach this was done by using multi-words. Finally, the combination of the within-word method with the two cross-word methods was tested. The word error rate (WER) measured for the baseline system was 12.75%. Compared to the baseline, a small but statistically significant improvement of 0.68% in W ER was measured for the within-word method, whereas both cross-word methods in isolation led to small, non-significant improvements. The combination of the within-word method and cross-word method 2 led to the best result: an absolute improvement of 1.12% in WER was found compared to the baseline, which is a relative improvement of 8.8% in WER.” Kessens, at Abstract. “In this research, we attempted to model two types of variation: within-word variation and cross-word variation. To this end, we used a general procedure in which pronunciation variation was modeled at the three different levels in the CSR: the lexicon, the phone models and the language model. We found that the best results were obtained when all of the steps of the general procedure were carried out, i.e. when pronunciation variants were incorporated at all three levels. Below, the results of incorporating pronunciation variants at all three levels are successively discussed.”

’993 Patent		
		Kessens, at 205. “In the first step, variants were only incorporated at the level of the *lexicon*. Compared to the baseline (SSS ^ MSS), an improvement was found for the within-word method and for the within-word method in combination with each of the two crossword methods. However, a deterioration was found for the two cross-word methods in isolation. A possible explanation for the deterioration for cross-word method 1 is related to the fact that the pronunciation variants of cross-word method 1 are very short (see Table 2); some of them consist of only one phone. Such short variants can easily be inserted; for instance, the plosives /k/ and /t/ might occasionally be inserted at places where clicks in the signal occur. Furthermore, this effect is facilitated by the high frequency of occurrence of the words involved, i.e. they are favored by the language model. Similar things might happen for cross-word method 2. Let us give an example to illustrate this: A possible variant of the multi-word “ik_wil” /IkwIl/ is /kwIl/. The latter might occasionally be confused with the word “wil” /wIl/. This confusion leads to a substitution, but effectively it is the insertion of the phone /k/. Consequently, insertion of /k/ and other phones is also possible in cross-word method 2, and this could explain the deterioration found for cross-word method 2.” Kessens, at 205. “When, in the second step, pronunciation variation is also incorporated at the level of the *phone models* (MSS ^ MMS), the CSR’s performance improved in all cases, except in the case of crossword method 2. A possible cause of this deterioration in performance could be that the phone models were not retrained properly. During forced recognition, the option for recognizing a pause between the separate parts of the multi-words was not given. As a consequence, if a pause occurred in the acoustic signal of a multi-word, the pause was used to train the surrounding phone models, which results in contaminated phone models. Error analysis revealed that in 5% of the cases a pause was indeed present within the multi-words in our training material. Further research will have to show whether this was the only cause of the deterioration in performance or whether there are other reasons why retraining phone models using multi-words did not lead to improvements.”

’993 Patent		
		Kessens, at 205. “In the third step, pronunciation variants were also incorporated at the level of the *language model* (MMS ^ MMM), which is beneficial to all methods. Moreover, the effect of adding variants to the language model is much larger for the crossword methods than for the within-word method. This is probably due to the fact that many recognition errors introduced in the first step (see above) are corrected when variants are also included in the language model. When cross-word variants are added to the lexicon (step 1), short sequences of only one or two phones long (like e.g. the phone /k/) can easily be inserted, as was argued above. The output of forced recognition reveals that the cross-word variants occur less frequently than the canonical pronunciations present in the baseline lexicon: on average in about 13% of the cases for cross-word method 1, and 9% for cross-word method 2. In the language model with cross-word variants included, the probability of these cross-word variants is thus lower than in the original language model and, consequently, it is most likely that they will be inserted less often.” Kessens, at 205. “One of the questions we posed in the introduction was what the best way of modeling crossword variation is. On the basis of our results we can conclude that when cross-word variation is modeled in isolation, cross-word method 2 performs better than cross-word method 1, but the difference is non-significant. In combination with the within-word method, cross-word method 2 leads to an improvement compared to the within-word method in isolation. This is not the case for cross-word method 1, which leads to a degradation in WER. Therefore, it seems that cross-word method 2 is more suitable for modeling cross-word pronunciation variation. It should be noted, however, that most of the improvements gained with cross-word method 2 are due to adding the multi-words to the lexicon and the language model. An explanation for these improvements is that by adding multi-words to the language model the span of the unigram and bigram increases for the most frequent word sequences in the training corpus. Thus, more context information can be used during the recognition process. Furthermore, it should also be noted that only a small amount of data was involved in the cross-word processes which were studied; only 6-9% of the words in the training corpus

’993 Patent		
		were affected by these processes. Therefore, we plan to test cross-word methods 1 and 2 for a larger amount of data and a larger number of cross-word processes.” Kessens, at 205-06. “To summarize, we obtained the best results when within-word pronunciation variation and cross-word pronunciation variation using multiword were modeled in combination, and all the steps of the general procedure had been carried out. Using only five phonological rules and 22 multi-words a relative improvement of 8.8% was found (12.75%-11.63%).” Kessens, at 207.

APPENDIX B-3
Invalidity Claim Chart for U.S. Patent No. 8,532,993 ('993 patent)
Helmer Strik, *Modeling pronunciation variation for ASR: A survey of the literature*, 1999 ("Strik")[1]

On October 16, 2020, Nuance narrowed the asserted '993 patent claims to 17 and 19. Nuance's Disclosure Limiting Asserted Claims of Phase 1 Patents. Claims 17 and 19 of the '993 patent are anticipated and/or rendered obvious by Strik alone or in combination with at least the following references. The chart below discloses how prior art references identified by Omilia disclose, either expressly or inherently, and/or render obvious each of the asserted claims:

(1) Steinbiss et al., *The Philips research system for large-vocabulary continuous-speech recognition*, Proc. of 3rd European Conference on Speech Communication and Technology EUROSPEECH '93, 2125-2128 (1993) ("Steinbiss")[2]

(2) Jain, et al., *Creating Speaker-Specific Phonetic Templates with a Speaker-Independent Phonetic Recognizer: Implications for Voice Dialing*, IEEE International Conference on Acoustics, Speech, and Signal Processing Conference Proceedings, 881-884 (1996) ("Jain")[3]

(3) Kessens et al., *Improving the performance of a Dutch CSR by modeling within-word and cross-word pronunciation variation*, Speech Communication 29 (1999) 193-207 ("Kessens")[4]

This invalidity claim chart is based on Omilia's present understanding of the asserted claims, the Court's August 6, 2020 Claim Construction Order (ECF. No. 241), Plaintiff ("Nuance's") initial and all subsequent supplemental Infringement Contentions, its Response to Omilia NLS' Supplemental Preliminary Non-Infringement and Invalidity Contentions, its Response to Omilia NLS' Interrogatory No. 9 and Omilia's investigation to date. Omilia is not adopting Nuance's apparent constructions, nor is Omilia admitting to the accuracy of any particular construction. Omilia reserves all rights to amend this invalidity claim chart if Nuance further amends its Infringement Contentions, or in response to any statements made by Nuance at any phase of this litigation. In addition, Discovery is still ongoing, and Omilia reserves its right to supplement its contentions with additional evidence as discovery continues.

The citations provided are exemplary and do not necessarily include each and every disclosure of the limitation in the references. Omilia has cited the most relevant portions of the identified prior art. Other portions of the identified prior art may additionally disclose, either

[1] Strik was published in 1999. Strik is prior art under at least pre-AIA 35 U.S.C. § 102(a), and (b).
[2] Steinbiss was published in September, 1993. Steinbiss is prior art under at least pre-AIA 35 U.S.C. § 102(a), and (b).
[3] Jain was published in 1996. Jain is prior art under at least pre-AIA 35 U.S.C. § 102(a), and (b).
[4] Kessens was published in 1999. Kessens is prior art under at least pre-AIA 35 U.S.C. § 102(a), and (b).

expressly or inherently, and/or render obvious one or more limitations of the asserted claims. To provide context or aid in understanding the prior art and the state of the art at the time of the alleged invention, Omilia reserves the right to rely on: (1) uncited portions of the identified prior art; (2) other prior art not identified herein; (3) references that show the state of the art (irrespective of whether such references themselves qualify as prior art to the asserted patents); (4) factual testimony from the inventors or authors of the prior art references, or purveyors of prior art devices; and/or (5) expert testimony.

Citations to a particular drawing or figure in the accompanying charts should be understood to include the description of the drawing or figure, as well as any text associated with the drawing or figure. Citations to particular text concerning a drawing or figure, should also be understood to encompass that drawing or figure as well. The identified prior art expressly and/or inherently discloses the features of the asserted claims under all proposed constructions of the asserted claims. To establish the inherency of certain features of the prior art to invalidate the asserted claims, Omilia may rely on cited or uncited portions of the prior art, other documents, factual testimony, and expert testimony. To the extent that the Court finds that this reference does not explicitly teach certain limitations in the asserted claims, such limitations would have been inherent and/or obvious and Omilia reserves the right to supplement or amend this claim chart to address such a finding.

’993 Patent		
Claim 17		
17.pre	A computer-readable storage device having instructions stored which, when executed on a processor, cause the processor to perform operations comprising:	To the extent that the preamble is limiting, Strik discloses, expressly or inherently, “computer-readable storage device having instructions stored which, when executed on a processor, cause the processor to perform operations:” In Strik, the process described was developed and implemented using an automatic speech recognition computer system using conventional computer elements. *See, e.g.*, “The focus in automatic speech recognition (ASR) research has gradually shifted from isolated words to conversational speech. Consequently, the amount of pronunciation variation present in the speech under study has gradually increased. Pronunciation variation will deteriorate the performance of an ASR system if it is not well accounted for. This is probably the main reason why research on modeling pronunciation variation for ASR has increased lately. In this contribution, we provide an overview of the publications on this topic, paying particular attention to the papers in

’993 Patent		
		this special issue and the papers presented at ‘the Rolduc workshop’. First, the most important characteristics that distinguish the various studies on pronunciation variation modeling are discussed. Subsequently, the issues of evaluation and comparison are addressed. Particular attention is paid to some of the most important factors that make it difficult to compare the different methods in an objective way. Finally, some conclusions are drawn as to the importance of objective evaluation and the way in which it could be carried out.” Strik, at Abstract. “It is obvious that each of these questions cannot be answered in isolation. On the contrary, the answers will be highly interdependent. Depending on the decision taken for each of the above questions, different methods for pronunciation variation modeling can be distinguished, as will appear from the following sections. For each question it is possible to identify a specific dimension along which a choice can be made. In this way a descriptive framework can be obtained to classify the various contributions to modeling pronunciation variation in ASR. Although it is certainly possible that the extremes of some of these dimensions will not occur in practice, this is irrelevant since their main function is to provide us with a framework for description.” Strik, at 229. “The majority of the contributions are concerned with variation at the segmental level. A common way of describing segmental pronunciation variation in the context of ASR is by indicating whether it refers to word-internal or to cross-word processes, because this choice is strongly related to the properties of the speech recognizer being used. As a matter of fact, the choice for word-internal variation, cross-word variation or both, is

’993 Patent		
		determined by factors such as the type of ASR, the language, and the level at which modeling will take place.” Strik, at 229. “Given that the recognition engines of most ASR systems consist of three components, there are three levels at which variation can be modeled: the lexicon, the acoustic models, and the language model. This is not to say that modeling at one level precludes modeling at one of the other levels; on the contrary, to obtain a good recognition system, it is necessary that concerted modeling happens on the three levels. Therefore, in most studies modeling takes place at more than one level. Nevertheless, in order to categorize the various studies, each category will be discussed separately in the following subsections.” Strik, at 233. “The question that arises at this point is: Is an objective evaluation and comparison of these methods at all possible? This question is not easy to answer. An obvious solution seems to be to use benchmark corpora and standard methods for evaluation (e.g., to give everyone the same canonical lexicon), like the NIST evaluations for automatic speech recognition and automatic speaker verification. This would solve a number of the problems mentioned above, but certainly not all of them. The most important problem that remains is the choice of the language. Like many other benchmark tests it could be (American) English. However, pronunciation variation and the ways in which it should be modeled can differ between languages, as argued above. Furthermore, for various reasons it would favor groups who do research on (American) English. Finally, using benchmarks would not solve the problem of differences between ASR systems.”

’993 Patent		
		Strik, at 240. “Finally, it is worth mentioning that at present most researchers use ‘standard ASR systems’ based on discrete segmental representations, HMMs to model the segments, and features that are computed per frame (usually cepstral features and their derivatives). Possibly, the underlying assumptions in these standard ASR systems are not optimal. One of the assumptions is that speech is made up of discrete segments, usually phone(me)s. Although this has long been one of the assumptions in linguistics too, the idea that speech can be phonologically represented as a sequence of discrete entities (the ‘absolute slicing hypothesis’, as formulated in (Goldsmith, 1976, pp. 16±17)) has proved to be untenable. In non-linear, autosegmental phonology (Goldsmith, 1976, 1990) an analysis has been proposed in which different features are placed on different tiers. The various tiers represent the parallel activities of the articulators in speech, which do not necessarily begin and end simultaneously. In turn the tiers are connected by association lines. In this way, it is possible to indicate that the mapping between tiers is not always one to one. Assimilation phenomena can then be represented by the spreading of one feature from one segment to the adjacent one. On the basis of this theory, Li Deng and his colleagues have built ASR systems with which promising results have been obtained (Deng and Sun, 1994).” Strik, at 242. In addition, it would be obvious to a person having ordinary skill in the art to combine Strik with Steinbiss. Both Strik and Steinbiss are in the same field of art. A person having ordinary skill in the art would be motivated to combine Strik with Steinbiss. A person having ordinary skill in the art considering Strik would be motivated to implement the system on a computer, as disclosed by Steinbiss. A person having ordinary skill in the art would have a reasonable expectation of success in implementing Strik

’993 Patent		
		on the computer of Steinbiss because the combination involves the predictable use of prior art elements according to their established functions. The motivation to combine these references would at least include: • Combining prior art elements according to known methods to yield predictable results (using a computer to perform speech recognition using computers was well known in the art); • Simple substitution of one known element for another to obtain predictable results (implementing the system of Strik on a computer, as disclosed by Steinbiss); • Use of known technique to improve similar devices (methods, or products) in the same way (the technique of using a computer to perform speech recognition was known)); • Applying a known technique to a known device (method, or product) ready for improvement to yield predictable results (using a computer to perform speech recognition using computers was well known in the art); • “Obvious to try”—choosing from a finite number of identified, predictable solutions, with a reasonable expectation of success (Strik teaches automatic speech recognition, which may be implemented using a computer); • Market forces and benefits associated with the known benefits of automatic speech recognition; and

’993 Patent		
		• Teaching of prior art would have lead a POSA to combine the references to arrive at a speech recognition system implemented on a computer. Steinbiss also discloses, expressly or inherently, “computer-readable storage device having instructions stored which, when executed on a processor, cause the processor to perform operations.” In Steinbiss, the process described was developed and implemented using PC based implementation. *See, e.g.*, “The system has been successfully applied to the American English DARPA RM task. Here, we report experimental results for a German 13 000-word Philips internal dictation task. In addition to the scientific prototype, a PC version has been set up which is described here for the first time.” Steinbiss, at 2125. “The organization of the paper is as follows. We first summarize the statistical approach to speech recognition and then describe the four main entities of our system: acoustic analysis, acoustic phonetic modelling, language modelling and search. A section with experiments on our internal dictation task follows. The final section describes a PC based implementation of our system.” Steinbiss, at 2125. “8. A PC Based Continuous Speech-Recognition System for Dictation . . . The system developed at Philips Dictation Systems, Vienna, and described here adopts this non-interactive approach and thus allows the person to

’993 Patent		
		dictate with a natural speaking style. After the speech is processed by the speech recognizer, the secretary has only to correct the recognition errors, which is both faster and a more interesting job to do.” Steinbiss, at 2128. “Speech recognition runs remotely on a PC which is connected to the network. An acoustic front-end performs the acoustic analysis. Recognition is sped up by a dedicated co-processor board containing application-specific ICs. Depending on the speaker and the specific boundary conditions, recognition with a 10 - 20 000-word vocabulary runs in 1 - 3 times real-time.” Steinbiss, at 2128.
17.a	approximating transcribed speech using a phonemic transcription dataset associated with a speaker, to yield a language model, where the phonemic transcription dataset is based on a pronunciation model of the speaker;	Strik discloses, expressly or inherently, the step of “approximating transcribed speech using a phonemic transcription dataset associated with a speaker, to yield a language model, where the phonemic transcription dataset is based on a pronunciation model of the speaker:” Strik discloses, expressly or inherently, the step of “***approximating transcribed speech using a phonemic transcription dataset*** associated with a speaker, to yield a language model, where the phonemic transcription dataset is based on a pronunciation model of the speaker:” In Strik, an ASR system is trained using automatically-segmented acoustic signals with pronunciation variation. *See, e.g.*, “Another feature that distinguishes the various approaches to modeling pronunciation variation in ASR is the source from which information on pronunciation variation is derived. In this connection, a distinction can be drawn between data-driven versus knowledge-based methods. The major difference between these two types of approaches is that in the former case

’993 Patent		
		the assumption is that the information on pronunciation variation has to be obtained in the first place. In knowledge-based approaches, on the other hand, it is assumed that this information is already available in the literature.” Strik, at 230. “The idea behind data-driven methods is that information on pronunciation variation has to be obtained directly from the signals (Bacchiani and Ostendorf, 1998, 1999; Blackburn and Young, 1995, 1996; Cremelie and Martens, 1995, 1997, 1998, 1999; Fosler-Lussier and Morgan, 1998, 1999; Fukada and Sagisaka, 1997; Fukada et al., 1998, 1999; Greenberg, 1998, 1999; Heine et al., 1998; Holmes and Russell, 1996; Holter, 1997; Holter and Svendsen, 1998, 1999; Imai et al., 1995; Mirghafori et al., 1995; Mokbel and Jouvet, 1998; Nock and Young, 1998; Peters and Stubley, 1998; Polzin and Waibel, 1998; Ravishankar and Eskenazi, 1997; Riley et al., 1998, 1999; Ristad and Yianilos, 1998; Sloboda and Waibel, 1996; Svendsen et al., 1995; Torre et al., 1997; Williams and Renals, 1998). To this end, the acoustic signals are analyzed in order to determine all possible ways in which the same word or phoneme is realized. A common stage in this analysis is transcribing the acoustic signals. Subsequently, the transcriptions can be used for different purposes, as will be explained in Sections 2.3 and 2.4. Transcriptions of the acoustic signals can be obtained either manually (Cremelie and Martens, 1995, 1997; Downey and Wiseman, 1997; Fosler-Lussier and Morgan, 1998, 1999; Greenberg, 1998, 1999; Heine et al., 1998; Mirghafori et al., 1995; Riley et al., 1998, 1999; Ristad and Yianilos, 1998; Wiseman and Downey, 1998) or (semi-) automatically (Adda-Decker and Lamel, 1998, 1999; Bacchiani and Ostendorf, 1998, 1999; Beulen et al., 1998; Cremelie and Martens, 1997, 1998, 1999; Fukada and Sagisaka, 1997; Fukada et al., 1998, 1999; Holter, 1997; Kessens and Wester, 1997; Kessens et al., 1999; Lehtinen and Safra, 1998; Mokbel and Jouvet, 1998; Ravishankar and Eskenazi, 1997; Riley et

’993 Patent		
		al., 1998, 1999; Schiel et al., 1998; Wester et al., 1998a; Svendsen et al., 1995; Torre et al., 1997; Williams and Renals, 1998). The latter is usually done either with a phone(me) recognizer (Fukada and Sagisaka, 1997; Fukada et al., 1998, 1999; Mokbel andJouvet, 1998; Ravishankar and Eskenazi, 1997; Torre et al., 1997; Williams and Renals, 1998) or by means of forced recognition (Adda-Decker and Lamel, 1998, 1999; Bacchiani and Ostendorf, 1998, 1999; Beulen et al., 1998; Cremelie and Martens, 1997, 1998, 1999; Kessens and Wester, 1997; Kessens et al., 1999; Lehtinen and Safra, 1998; Riley et al., 1998, 1999; Schiel et al., 1998; Wester et al., 1998a).” Strik, at 230. “Before variants can be selected, they have to be obtained, in the first place. Sometimes the pronunciation variants are generated manually (Aubert and Dugast, 1995; Riley et al., 1998, 1999) or selected from enumerated lists (Flach, 1995), but usually they are generated automatically by means of various procedures: • rules (Adda-Decker and Lamel, 1998, 1999; Aubert and Dugast, 1995; Cohen and Mercer, 1975; Cremelie and Martens, 1995, 1997, 1998, 1999; Flach, 1995; Kessens and , 1997; Kessens et al., 1999; Mercer and Cohen, 1987; Nock and Young, 1998; Ravishankar and Eskenazi, 1997; Schiel et al., 1998; et al., 1998a), • ANNs (Fukada and Sagisaka, 1997; Fukada et al., 1998, 1999), • grapheme-to-phoneme converters (Lehtinen and Safra, 1998), • phone(me) recognizers (Mokbel and Jouvet, 1998; Nock and Young, 1998; Ravishankar and Eskenazi, 1997; Sloboda and Waibel, 1996; Williams and Renals, 1998), • optimization with maximum likelihood criterion (Holter, 1997; Holter and Svendsen, 1998, 1999) and

’993 Patent		
		• decision trees (Fosler-Lussier and Morgan, 1998, 1999; Riley et al., 1998, 1999). Strik, at 234. Strik discloses, expressly or inherently, the step of “approximating transcribed speech using ***a phonemic transcription dataset associated with a speaker***, to yield a language model, ***where the phonemic transcription dataset is based on a pronunciation model of the speaker***.” In Strik, an ASR system is trained using automatically-segmented acoustic signals with pronunciation variation A POSA would further understand that the acoustic signals are associated with a particular speaker. A POSA would further understand that the process of segmenting acoustic signals with pronunciation variation disclosed by Strik is “based on a pronunciation model of the speaker.” *See, e.g.*, “In forced recognition (which is also called forced alignment), the ASR system can only choose between the pronunciation variants of a word, and not between all words present in the lexicon, as is the case for a ‘normal’ ASR system. Consequently, forced recognition can be employed to decide which pronunciation variant best matches the signal, and in this way a new transcription can be obtained (see also Section 2.4.2). The performance of forced recognition has been evaluated in (Wester et al., 1998a,b, 1999), by comparing it with the performance of humans that carried out the same tasks. It turned out that for the tasks studied in (Wester et al., 1998a,b, 1999), the performance of the human listeners and forced recognition were similar, and that on average, the degree of agreement between ASR system and listeners is only slightly lower than that between listeners. Therefore, forced recognition seems to be a suitable tool for obtaining information on pronunciation (variation). However, since in (Wester et al., 1999) it is

’993 Patent		
		shown that the agreement depends on the properties of the ASR system used, one should be cautious in applying such a tool.” Strik, at 230. “Although many studies contained in this issue are not completely data-driven or knowledge-based, it is generally possible to say whether the starting point of the research was mainly data-driven or knowledge-based (see the references above). However, most of them cannot be said to be completely bottom-up or top-down, because in none of these studies is the direction of the developing process solely upward or downward, but the flow of information is in both directions. For example, in many data-driven studies the results of the bottom-up analyses are used to change the lexicon and the altered lexicon is then used during recognition in a top-down manner. Similarly, knowledge-based methods are usually not strictly top-down, e.g. because in many of them the rules applied to generate pronunciation variants may be altered on the basis of information derived from analysis of the acoustic signals.” Strik, at 231. “The obvious alternative to using formalizations is to use information that is not formalized, but enumerated. Again, this can be done either in a data-driven or in a knowledge-based manner. In data-driven studies, the bottom-up transcriptions can be used to list all pronunciation variants of one and the same word. These variants and their transcriptions (or a selection of them) can then be added to the lexicon. Alternatively, in knowledge-based studies it is possible to add all the variants of one and the same word contained in a pronunciation dictionary. Quite clearly, when no formalization is used, it is not necessary to generate the variants because they are already available.”

'993 Patent		
		Strik, at 232. "Given that the recognition engines of most ASR systems consist of three components, there are three levels at which variation can be modeled: the lexicon, the acoustic models, and the language model. This is not to say that modeling at one level precludes modeling at one of the other levels; on the contrary, to obtain a good recognition system, it is necessary that concerted modeling happens on the three levels. Therefore, in most studies modeling takes place at more than one level. Nevertheless, in order to categorize the various studies, each category will be discussed separately in the following subsections." Strik, at 233. "Before variants can be selected, they have to be obtained, in the first place. Sometimes the pronunciation variants are generated manually (Aubert and Dugast, 1995; Riley et al., 1998, 1999) or selected from enumerated lists (Flach, 1995), but usually they are generated automatically by means of various procedures: • rules (Adda-Decker and Lamel, 1998, 1999; Aubert and Dugast, 1995; Cohen and Mercer, 1975; Cremelie and Martens, 1995, 1997, 1998, 1999; Flach, 1995; Kessens and Wester, 1997; Kessens et al., 1999; Mercer and Cohen, 1987; Nock and Young, 1998; Ravishankar and Eskenazi, 1997; Schiel et al., 1998; Wester et al., 1998a), • ANNs (Fukada and Sagisaka, 1997; Fukada et al., 1998, 1999), • grapheme-to-phoneme converters (Lehtinen and Safra, 1998), • phone(me) recognizers (Mokbel and Jouvet, 1998; Nock and Young, 1998; Ravishankar and Eskenazi, 1997; Sloboda and Waibel, 1996; Williams and Renals, 1998),

’993 Patent		
		• optimization with maximum likelihood criterion (Holter, 1997; Holter and Svendsen, 1998, 1999) and • decision trees (Fosler-Lussier and Morgan, 1998, 1999; Riley et al., 1998, 1999). Strik, at 234. “Second, the fact that a method in which variants are taken directly from transcriptions of the acoustic signals works better than a rule-based one could also be due to the particular nature of the rules in question. As was pointed out in Section 2.2, rules taken from the literature are not always the optimal ones to model variation in spontaneous speech, while information obtained from data may be much better suited for this purpose.” Strik, at 234. In addition, it would be obvious to a person having ordinary skill in the art to combine Strik with Jain. Both Strik and Jain are in the same field of art. A person having ordinary skill in the art would be motivated to combine Strik with Jain. A person having ordinary skill in the art considering Strik’s disclosure that variants “usually they are generated automatically by means of various procedures” would be motivated to seek out a system that automatically captures a phonetic representation of pronunciation variants, as disclosed by Jain. A person having ordinary skill in the art would have a reasonable expectation of success in implementing Strik using the phonetic transcriptions provided by Jain because the combination involves the predictable use of prior art elements according to their established functions. The motivation to combine these references would at least include:

’993 Patent		
		• Combining prior art elements according to known methods to yield predictable results (using speaker-dependent training data to improve accuracy for a user); • Simple substitution of one known element for another to obtain predictable results (adding additional pronunciation variants to improve accuracy of a speech recognition system); • Use of known technique to improve similar devices (methods, or products) in the same way (using speaker-dependent pronunciation data to improve the accuracy of a speech recognition system)); • Applying a known technique to a known device (method, or product) ready for improvement to yield predictable results (using speaker-dependent pronunciation was well known in the art); • “Obvious to try”—choosing from a finite number of identified, predictable solutions, with a reasonable expectation of success (there are a finite number of known techniques to improve accuracy for speech recognition systems); • Market forces and benefits associated with the known benefits of automatic speech recognition; and • Teaching of prior art would have lead a POSA to combine the references to arrive at a system that includes transcription data reflecting pronunciation data from a speaker. Jain also discloses, expressly or inherently, the step of “***approximating transcribed speech using a phonemic transcription dataset associated with a speaker***, to yield a language model, ***where the phonemic transcription dataset is based on a pronunciation model of the speaker***:”

’993 Patent		
		“In this section, we describe the method for generating the speaker-specific models. Our speaker-specific models represent utterances as sequences of phonemes rather than acoustic parameters. A speaker-independent phonetic front end is used to generate the string of phonemes. For this approach to work, all that is needed is that the phonetic frontend behave in a consistent manner, i.e. it must produce the same (or nearly the same) string of phonemes each time the word is spoken.” In Jain, speaker-specific models are used to transcribe telephone speech into its phonetic subparts. A POSA would further understand that the “telephone speech” is associated with a particular speaker. *See, e.g.*, Jain, at 881. “The following steps produce a phonetic transcription: 1. Data Capture: Telephone speech is sampled at 8 kHz. Unnecessary silence at the beginning and end of the utterance is removed by end-point detection. 2. Feature Extraction: Seventh order RASTA features [2] are computed for every 10ms of speech using a 10ms window. This yields eight coefficients per frame.” Jain, at 881. “Another method is to force-align each phoneme string (which was generated using the template generation technique) with all the other waveforms corresponding to the same label and compute the average Viterbi score. The template with the maximum average score is selected as the word-model for that label. Table 1 shows the error rates obtained with forced-alignment. It can be seen that the error rates decreased substantially for all the 4 channels.

’993 Patent

Telephone Channel	1	2	3	4
Baseline	7.9	12.0	8.7	18.5
Forced Alignment	4.8	8.0	5.8	15.5

Table 1. Effect of Forced Alignment ”

Jain, at 882-83.

“It is well known that speaker dependent systems perform better than speaker-independent systems. In several commercial products, speaker adaptation is used to adapt speaker-independent models to the current user.”

Jain at 883.

Strik discloses, expressly or inherently, the step of “approximating transcribed speech using a phonemic transcription dataset associated with a speaker, ***to yield a language model***, where the phonemic transcription dataset is based on a pronunciation model of the speaker.” In Strik, a language model is created using pronunciation variants. *See, e.g.*,

“Another component in which pronunciation variation can be taken into account is the language model (LM) (Cremelie and Martens, 1995, 1997, 1998, 1999; Deshmukh et al., 1996; Finke and Waibel, 1997; Fukada et al., 1998, 1999; Kessens et al., 1999; Lehtinen and Safra, 1998; Perennou and Brieussel-Pousse, 1998; Pousse and Perennou, 1997; Schiel et al., 1998; Wester et al., 1998a; Zeppenfeld et al., 1997). This can be done in several ways, as will be discussed below.”

Strik, at 236.

“*Method 1*. The easiest solution is to simply add the variants to the lexicon, and not to change the LMs at all. In this case, for every variant the

’993 Patent		
		probabilities for the word it belongs to are used. Since the statistics for the variants are not used, it is obvious that this is a sub-optimal solution. In the following two methods the statistics for the variants are employed.” Strik, at 236. “*Method 2*. The second solution is to use the variants themselves (instead of the underlying words) to calculate the N-grams (Kessens et al., 1999; Schiel et al., 1998; Strik et al., 1998a). For this procedure, a transcribed corpus is needed which contains information about the realized pronunciation variants. These transcriptions can be obtained in various ways, as has been discussed in Sections 2.2 and 2.4. The goal of this method is to find the string of variants V which maximizes $P(X\|V) * P(V)$.” Strik, at 236-37. “Another important difference between the two methods is that in the third method the context-dependence of pronunciation variants is not modeled directly in the LM. This can be a disadvantage as pronunciation variation is often context-dependent, e.g., liaison in French (Perennou and Brieussel-Pousse, 1998; Pousse and Perennou, 1997). Within the third method, this deficiency can be overcome by using classes of words instead of the words themselves, i.e., the classes of words that do or do not allow liaison (Perennou and Brieussel-Pousse, 1998; Pousse and Perennou, 1997). The probability of a pronunciation variant for a certain class is then represented in $P(V\|W)$, while the probability of sequences of word classes is stored in $P(W)$.” Strik, at 237. “One of the most common ways of modeling pronunciation variation is to add pronunciation variants to the lexicon (see Section 2.4.1). This method

’993 Patent		
		can be applied fairly easily and it appears to improve recognition performance. However, a problem with this approach is that certain words have numerous variants with very different frequencies of occurrence. Some quantitative data on this phenomenon can be found in Table 2 on page 50 of Greenberg (1998). For instance, if we look at the data for ‘that’, we can see that this word appears 328 times in the corpus used by Greenberg, that it has 117 different pronunciations and that the single most common variant only covers 11% of the pronunciations. The coverage of the other variants will probably decrease gradually from 11% to almost zero. In principle one could include all 117 variants in the lexicon and it is possible that this will improve recognition of the word ‘that’. However, this is also likely to increase confusability. If many variants of a large number of words are included in the lexicon the confusability can increase to such an extent that recognition performance may eventually decrease. This implies that variant selection constitutes an essential part of this approach.” Strik, at 240-41.
17.b	incorporating, into the language model, pronunciation probabilities associated with respective unique labels for each different pronunciation of a word, wherein the respective unique label for a most frequent word indicates a special status in the language model; and	Strik discloses, expressly or inherently, the step of “incorporating, into the language model, pronunciation probabilities associated with respective unique labels for each different pronunciation of a word, wherein the respective unique label for a most frequent word indicates a special status in the language model.” Strik discloses, expressly or inherently, the step of “***incorporating, into the language model, pronunciation probabilities associated with respective unique labels for each different pronunciation of a word***, wherein the respective unique label for a most frequent word indicates a special status in the language model.” In Strik, pronunciation variants for

’993 Patent		
		words and their probabilities first associated with unique labels, which are then incorporated into a language model. *See, e.g.*, “Another feature that distinguishes the various approaches to modeling pronunciation variation in ASR is the source from which information on pronunciation variation is derived. In this connection, a distinction can be drawn between data-driven versus knowledge-based methods. The major difference between these two types of approaches is that in the former case the assumption is that the information on pronunciation variation has to be obtained in the first place. In knowledge-based approaches, on the other hand, it is assumed that this information is already available in the literature.” Strik, at 230. “Given that the recognition engines of most ASR systems consist of three components, there are three levels at which variation can be modeled: the lexicon, the acoustic models, and the language model. This is not to say that modeling at one level precludes modeling at one of the other levels; on the contrary, to obtain a good recognition system, it is necessary that concerted modeling happens on the three levels. Therefore, in most studies modeling takes place at more than one level. Nevertheless, in order to categorize the various studies, each category will be discussed separately in the following subsections.” Strik, at 233. “At the level of the lexicon, pronunciation variation is usually modeled by adding pronunciation variants (and their transcriptions) to the lexicon (Adda-Decker and Lamel, 1998, 1999; Aubert and Dugast, 1995; Beulen et al., 1998; Bonaventura et al., 1998; Cohen and Mercer, 1975; Cremelie and Martens, 1995, 1997, 1998, 1999; Downey and Wiseman, 1997;

’993 Patent		
		Ferreiros et al., 1998; Finke and Waibel, 1997; Fukada et al., 1998, 1999; Holter, 1997; Holter and Svendsen, 1998, 1999; Imai et al., 1995; Kessens and Wester, 1997; Kessens et al., 1999; Lamel and Adda, 1996; Lehtinen and Safra, 1998; Mercer and Cohen, 1987; Mokbel and Jouvet, 1998; Nock and Young, 1998; Ravishankar and Eskenazi, 1997; Riley et al., 1998, 1999; Roach and Arnfield, 1998; Sloboda and Waibel, 1996; Torre et al., 1997; Wester et al., 1998a; Williams and Renals, 1998; Wiseman and Downey, 1998; Zeppenfeld et al., 1997). The rationale behind this procedure is that with multiple transcriptions of the same word the chance is increased that for an incoming signal the speech recognizer selects a transcription belonging to the correct word. In turn, this should lead to lower error rates.” Strik, at 233. “The obvious alternative to using formalizations is to use information that is not formalized, but enumerated. Again, this can be done either in a data-driven or in a knowledge-based manner. In data-driven studies, the bottom-up transcriptions can be used to list all pronunciation variants of one and the same word. These variants and their transcriptions (or a selection of them) can then be added to the lexicon. Alternatively, in knowledge-based studies it is possible to add all the variants of one and the same word contained in a pronunciation dictionary. Quite clearly, when no formalization is used, it is not necessary to generate the variants because they are already available.” Strik, at 232. “At the level of the lexicon, pronunciation variation is usually modeled by adding pronunciation variants (and their transcriptions) to the lexicon (Adda-Decker and Lamel, 1998, 1999; Aubert and Dugast, 1995; Beulen et al., 1998; Bonaventura et al., 1998; Cohen and Mercer, 1975; Cremelie

’993 Patent		
		and Martens, 1995, 1997, 1998, 1999; Downey and Wiseman, 1997; Ferreiros et al., 1998; Finke and Waibel, 1997; Fukada et al., 1998, 1999; Holter, 1997; Holter and Svendsen, 1998, 1999; Imai et al., 1995; Kessens and Wester, 1997; Kessens et al., 1999; Lamel and Adda, 1996; Lehtinen and Safra, 1998; Mercer and Cohen, 1987; Mokbel and Jouvet, 1998; Nock and Young, 1998; Ravishankar and Eskenazi, 1997; Riley et al., 1998, 1999; Roach and Arnfield, 1998; Sloboda and Waibel, 1996; Torre et al., 1997; Wester et al., 1998a; Williams and Renals, 1998; Wiseman and Downey, 1998; Zeppenfeld et al., 1997). The rationale behind this procedure is that with multiple transcriptions of the same word the chance is increased that for an incoming signal the speech recognizer selects a transcription belonging to the correct word. In turn, this should lead to lower error rates.” Strik, at 233. “However, adding pronunciation variants to the lexicon usually also introduces new errors because the acoustic confusability within the lexicon increases, i.e., the added variants can be confused with those of other entries in the lexicon. This can be minimized by making an appropriate selection of the pronunciation variants, by, for instance, adding only the set of variants for which the balance between solving old errors and introducing new ones is positive. Therefore, in many studies tests are carried out to determine which set of pronunciation variants leads to the largest gain in performance (Cremelie and Martens, 1995, 1997, 1998, 1999; Fukada et al., 1998, 1999; Holter, 1997; Holter and Svendsen, 1998, 1999; Imai et al., 1995; Kessens and Wester, 1997; Kessens et al., 1999; Lehtinen and Safra, 1998; Mokbel and Jouvet, 1998; Nock and Young, 1998; Riley et al., 1998, 1999; Sloboda and Waibel, 1996; Torre et al., 1997; Wester et al., 1998a). For this purpose, different criteria can be used, such as:

’993 Patent		
		• frequency of occurrence of the variants (Kessens and Wester, 1997; Kessens et al., 1999; Ravishankar and Eskenazi, 1997; Riley et al., 1998, 1999; Schiel et al., 1998; Wester et al., 1998a; Williams and Renals, 1998), • a maximum likelihood criterion (Holter, 1997; Holter and Svendsen, 1998, 1999; Imai et al., 1995), • confidence measures (Sloboda and Waibel, 1996), and • the degree of confusability between the variants (Sloboda and Waibel, 1996; Torre et al., 1997). A description of a method to detect confusable pairs of words or transcriptions is also given in (Roe and Riley, 1994). If rules are used to generate pronunciation variants, then certain rules can be selected (and others discarded), as in (Cremelie and Martens, 1995, 1997, 1998, 1999; Lehtinen and Safra, 1998; Schiel et al., 1998) where rules are selected on the basis of their frequency and application likelihood.” Strik, at 233-34. “As was mentioned earlier, multi-words can also be added to the lexicon, in an attempt to model cross-word variation at the level of the lexicon. Optionally, the pronunciation variants of multi-words can also be included in the lexicon. By using multi-words Beulen et al. (1998) and Wester et al. (Kessens et al., 1999; Wester et al., 1998a) achieve a substantial improvement, while for Nock and Young (1998) this was not the case.” Strik, at 234. “An obvious way of optimizing the acoustic models is by using forced recognition. In Section 2.2 we already explained how forced recognition can be employed to calculate new transcriptions of the signals. In turn, the new transcriptions can be used to train new acoustic models. These new

’993 Patent		
		acoustic models can then be used to do forced recognition again, etc. In other words, this process can be iterated. We will refer to this procedure as iterative transcribing. Forced recognition and iterative transcribing have been used often to obtain improved transcriptions and improved acoustic models (Aubert and Dugast, 1995; Bacchiani and Ostendorf, 1998; Bacchiani and Ostendorf, 1999; Beulen et al., 1998; Finke and Waibel, 1997; Kessens and Wester, 1997; Kessens et al., 1999; Riley et al., 1998, 1999; Schiel et al., 1998; Sloboda and Waibel, 1996; Wester et al., 1998a).” Strik, at 235. “In forced recognition, pronunciation variants are present in the lexicon during training in order to train new acoustic models. Optionally, pronunciation variants can be retained in the lexicon during recognition (testing). In general, using the variants during recognition is more beneficial than using variants during training, while the best results are obtained when multiple variants are included during both training and recognition (Kessens and Wester, 1997; Kessens et al., 1999; Lamel and Adda, 1996; Wester et al., 1998a). Therefore, it seems worthwhile to test the procedure of forced recognition because it is a relatively straightforward procedure that can be applied almost completely automatically and because it usually gives an improvement over and above that of using multiple variants during recognition only.” Strik, at 235. “Another component in which pronunciation variation can be taken into account is the language model (LM) (Cremelie and Martens, 1995, 1997, 1998, 1999; Deshmukh et al., 1996; Finke and Waibel, 1997; Fukada et al., 1998, 1999; Kessens et al., 1999; Lehtinen and Safra, 1998; Perennou and Brieussel-Pousse, 1998; Pousse and Perennou, 1997; Schiel et al.,

’993 Patent		
		1998; Wester et al., 1998a; Zeppenfeld et al., 1997). This can be done in several ways, as will be discussed below.” Strik, at 236. “*Method 1*. The easiest solution is to simply add the variants to the lexicon, and not to change the LMs at all. In this case, for every variant the probabilities for the word it belongs to are used. Since the statistics for the variants are not used, it is obvious that this is a sub-optimal solution. In the following two methods the statistics for the variants are employed.” Strik, at 236. “*Method 2*. The second solution is to use the variants themselves (instead of the underlying words) to calculate the N-grams (Kessens et al., 1999; Schiel et al., 1998; Wester et al., 1998a). For this procedure, a transcribed corpus is needed which contains information about the realized pronunciation variants. These transcriptions can be obtained in various ways, as has been discussed in Sections 2.2 and 2.4. The goal of this method is to find the string of variants V which maximizes $P(X\|V) * P(V)$.” Strik, at 236-37. “Another important difference between the two methods is that in the third method the context-dependence of pronunciation variants is not modeled directly in the LM. This can be a disadvantage as pronunciation variation is often context-dependent, e.g., liaison in French (Perennou and Brieussel-Pousse, 1998; Pousse and Perennou, 1997). Within the third method, this deficiency can be overcome by using classes of words instead of the words themselves, i.e., the classes of words that do or do not allow liaison (Perennou and Brieussel-Pousse, 1998; Pousse and Perennou, 1997). The probability of a pronunciation variant for a certain class is then

’993 Patent		
		represented in $P(V\|W)$, while the probability of sequences of word classes is stored in $P(W)$. Strik, at 237. “In trying to draw general conclusions as to the effectiveness of these methods, one is tempted to conclude that the method for which the largest improvement was observed is the best one. In this respect some comment is in order. First, it is unlikely that there will be one single best approach, as the tasks of the various systems are very different. Second, we are not interested in finding a winner, but in understanding how pronunciation variation can best be approached. So, even a method that does not produce any significant reduction in WER may turn out to be extremely valuable because it increases our understanding of pronunciation variation. Third, it is wrong to take the change in WER as the only criterion for evaluation, because this change is dependent on at least three different factors: (1) the corpora, (2) the ASR system and (3) the baseline system. This means that improvements in WER can be compared with each other only if in the methods under study these three elements were identical or at least similar. It is obvious that in the majority of the methods presented these three elements are not kept constant, but are usually very different. In the following sections we discuss these differences and try to explain why this makes it difficult to compare the various methods and, in particular, the results obtained with each of them.” Strik, at 238. Strik discloses, expressly or inherently, the step of “incorporating, into the language model, pronunciation probabilities associated with respective unique labels for each different pronunciation of a word, ***wherein the respective unique label for a most frequent word indicates a special status in the language model***.” In Strik, pronunciation variants and their

’993 Patent		
		probabilities result from a “variant selection” process, which selects the most frequent words and identifies a special status. *See, e.g.*, “One of the most common ways of modeling pronunciation variation is to add pronunciation variants to the lexicon (see Section 2.4.1). This method can be applied fairly easily and it appears to improve recognition performance. However, a problem with this approach is that certain words have numerous variants with very different frequencies of occurrence. Some quantitative data on this phenomenon can be found in Table 2 on page 50 of Greenberg (1998). For instance, if we look at the data for ‘that’, we can see that this word appears 328 times in the corpus used by Greenberg, that it has 117 different pronunciations and that the single most common variant only covers 11% of the pronunciations. The coverage of the other variants will probably decrease gradually from 11% to almost zero. In principle one could include all 117 variants in the lexicon and it is possible that this will improve recognition of the word ‘that’. However, this is also likely to increase confusability. If many variants of a large number of words are included in the lexicon the confusability can increase to such an extent that recognition performance may eventually decrease. This implies that variant selection constitutes an essential part of this approach.” Strik, at 240-41. “An obvious criterion for variant selection is frequency of occurrence. Adding very frequent variants is likely to produce a more substantial improvement than adding infrequent variants. However, there is no clear distinction between frequent and infrequent pronunciation variants. Furthermore, besides frequency of occurrence, there will be other important factors that influence recognition performance. For instance, some pronunciation variants will probably constitute no problem for the ASR system, in the sense that they will be recognized correctly even

’993 Patent		
		though they (slightly) differ from the representation stored in the lexicon. Other spoken variants will probably cause frequent errors during recognition. In order to improve the performance of the ASR system, it is necessary to know which variants cause recognition errors (and which do not). Furthermore, adding pronunciation variants to the lexicon can solve some recognition errors, but it will certainly also introduce new ones. To optimize performance one should add those variants for which the balance between solving old errors and introducing new errors is positive. Confusability during the decoding process is a central issue in this respect. However, it will be difficult to predict a priori what the confusability during decoding will be. A manner in which our insight on this topic could be enhanced, is by doing error analysis, as will be discussed below.” Strik, at 241. In addition, it would be obvious to a person having ordinary skill in the art to combine Strik with Kessens. Both Strik and Kessens are in the same field of art. Strik discloses that it is beneficial to include pronunciation variants and a potential issue is that this may increase “confusability.” Kessens teaches techniques that include pronunciation variants and protect against performance degradation caused by the increase in “confusability.” A person having ordinary skill in the art would therefore be motivated to use the disclosures of Kessens to implement the system of Wester. A person having ordinary skill in the art would have a reasonable expectation of success in implementing Wester using the disclosure of Kessens because the combination involves the predictable use of prior art elements according to their established functions. The motivation to combine these references would at least include:

’993 Patent		
		• Combining prior art elements according to known methods to yield predictable results (including pronunciation variants in a lexicon to improve performance of a speech recognizer); • Simple substitution of one known element for another to obtain predictable results (including pronunciation variants in a lexicon was well known in the art); • Use of known technique to improve similar devices (methods, or products) in the same way (using pronunciation variants and limiting their use was known)); • Applying a known technique to a known device (method, or product) ready for improvement to yield predictable results (using data-driven techniques to create a language model for speech recognition was well known in the art); • “Obvious to try”—choosing from a finite number of identified, predictable solutions, with a reasonable expectation of success (there are a finite number of ways to incorporate variant pronunciations in a speech recognition system); • Market forces and benefits associated with the known benefits of automatic speech recognition; and • Teaching of prior art would have lead a POSA to combine the references to arrive at a language model for speech recognition. Kessens also discloses, expressly or inherently, the step of “incorporating, into the language model, pronunciation probabilities associated with respective unique labels for each different pronunciation of a word, ***wherein the respective unique label for a most frequent word indicates a special status in the language model***.” In Kessens, the 50 most frequently

’993 Patent		
		occurring word sequences are selected, and given a special status. *See, e.g.*, “The first step in cross-word method 1 consisted of selecting the 50 most frequently occurring word sequences from our training material. Next, from those 50 word sequences we chose those words which are sensitive to the cross-word processes cliticization, contraction and reduction. This led to the selection of seven words which made up 9% of all the words in the training corpus (see Table 2). The variants of these words were added to the lexicon and the rest of the steps of the general procedure were carried out (see Section 2.2). Table 2 shows the selected words (column 1), the total number of times the word occurs in the training material (column 2), their baseline transcriptions (column 3) and their added cross-word variants (column 4).” Kessens, at 199. “For all methods, the best results are obtained when pronunciation variants are used during training and recognition, and when they are added to the language model (MMM). All methods lead to an improvement in the CSR’s performance when their results are compared to the result of the baseline (SSS). These improvements are summed up in Table 5. Modeling within-word variation in isolation gives a significant improvement of 0.68%, and in combination with cross-word method 2, the improvement is also significant.” Kessens, at 204.
17.c	after incorporating the pronunciation probabilities into the language model, recognizing an utterance using the language model.	Strik discloses, expressly or inherently, the step of “after incorporating the pronunciation probabilities into the language model, recognizing an utterance using the language model:” Strik discloses testing the recognizer

’993 Patent		
		with the language model as well as the resulting performance disclosed. *See, e.g.*, “The focus in automatic speech recognition (ASR) research has gradually shifted from isolated words to conversational speech. Consequently, the amount of pronunciation variation present in the speech under study has gradually increased. Pronunciation variation will deteriorate the performance of an ASR system if it is not well accounted for. This is probably the main reason why research on modeling pronunciation variation for ASR has increased lately. In this contribution, we provide an overview of the publications on this topic, paying particular attention to the papers in this special issue and the papers presented at ‘the Rolduc workshop’. First, the most important characteristics that distinguish the various studies on pronunciation variation modeling are discussed. Subsequently, the issues of evaluation and comparison are addressed. Particular attention is paid to some of the most important factors that make it difficult to compare the different methods in an objective way. Finally, some conclusions are drawn as to the importance of objective evaluation and the way in which it could be carried out.” Strik, at Abstract. “The question that arises at this point is: Is an objective evaluation and comparison of these methods at all possible? This question is not easy to answer. An obvious solution seems to be to use benchmark corpora and standard methods for evaluation (e.g., to give everyone the same canonical lexicon), like the NIST evaluations for automatic speech recognition and automatic speaker verification. This would solve a number of the problems mentioned above, but certainly not all of them. The most important problem that remains is the choice of the language. Like many other benchmark tests it could be (American) English. However, pronunciation variation and the ways in which it should be modeled can differ between languages, as argued

’993 Patent		
		above. Furthermore, for various reasons it would favor groups who do research on (American) English. Finally, using benchmarks would not solve the problem of differences between ASR systems.” Strik, at 240. “Finally, it is worth mentioning that at present most researchers use ‘standard ASR systems’ based on discrete segmental representations, HMMs to model the segments, and features that are computed per frame (usually cepstral features and their derivatives). Possibly, the underlying assumptions in these standard ASR systems are not optimal. One of the assumptions is that speech is made up of discrete segments, usually phone(me)s. Although this has long been one of the assumptions in linguistics too, the idea that speech can be phonologically represented as a sequence of discrete entities (the ‘absolute slicing hypothesis’, as formulated in (Goldsmith, 1976, pp. 16-17)) has proved to be untenable. In non-linear, autosegmental phonology (Goldsmith, 1976, 1990) an analysis has been proposed in which different features are placed on different tiers. The various tiers represent the parallel activities of the articulators in speech, which do not necessarily begin and end simultaneously. In turn the tiers are connected by association lines. In this way, it is possible to indicate that the mapping between tiers is not always one to one. Assimilation phenomena can then be represented by the spreading of one feature from one segment to the adjacent one. On the basis of this theory, Li Deng and his colleagues have built ASR systems with which promising results have been obtained (Deng and Sun, 1994).” Strik, at 242. “One of the most common ways of modeling pronunciation variation is to add pronunciation variants to the lexicon (see Section 2.4.1). This method can be applied fairly easily and it appears to improve recognition

’993 Patent		
		performance. However, a problem with this approach is that certain words have numerous variants with very different frequencies of occurrence. Some quantitative data on this phenomenon can be found in Table 2 on page 50 of Greenberg (1998). For instance, if we look at the data for ‘that’, we can see that this word appears 328 times in the corpus used by Greenberg, that it has 117 different pronunciations and that the single most common variant only covers 11% of the pronunciations. The coverage of the other variants will probably decrease gradually from 11% to almost zero. In principle one could include all 117 variants in the lexicon and it is possible that this will improve recognition of the word ‘that’. However, this is also likely to increase confusability. If many variants of a large number of words are included in the lexicon the confusability can increase to such an extent that recognition performance may eventually decrease. This implies that variant selection constitutes an essential part of this approach.” Strik, at 240-41. “In most studies mentioned above the emphasis was on reduction of the error rates. However, the difference in the error rates of two systems is only a global measure which does not provide information about the details of the differences in the recognition results. Consequently, in most studies it is not possible to find out how and why improvements were obtained. In order to do so the recognition errors should be studied in more detail, i.e., more detailed error analysis should be carried out. This can be done by comparing the errors in the recognition results between the old system and the new one. In addition, error analysis could be used not only post hoc, to test the effect of a specific method, but also before applying the method. For instance, it would be informative to know beforehand how many and what kind of errors are made so as to be able to estimate the

’993 Patent		
		maximum amount of improvement that can be achieved. In turn this could constitute a criterion in deciding whether to test the method at all.” Strik, at 241. “At this point, it may be useful to try to make a general assessment of the state of the art in research on modeling pronunciation variation for ASR. For example, we could start by relating the results obtained so far to the expectations researchers had at the beginning. It is difficult, though, to estimate the researchers’ expectations about the gain in recognition performance that could be obtained by modeling pronunciation variation. In any case, judging by the effort that has gone in this type of research one could conclude that there were high expectations. The results reported so far vary from 0% to 20% relative reduction in the WER. These findings can be interpreted either positively or negatively. Positively: modeling pronunciation variation often improves recognition performance, sometimes even by 20%. Negatively: sometimes recognition performance increases by about 20%, but in most cases improvements are marginal. At the Rolduc workshop the general feeling seemed to be that the results obtained so far did not live up to the expectations.” Strik, at 241-42. “Finally, it is worth mentioning that at present most researchers use ‘standard ASR systems’ based on discrete segmental representations, HMMs to model the segments, and features that are computed per frame (usually cepstral features and their derivatives). Possibly, the underlying assumptions in these standard ASR systems are not optimal. One of the assumptions is that speech is made up of discrete segments, usually phone(me)s. Although this has long been one of the assumptions in linguistics too, the idea that speech can be phonologically represented as a sequence of discrete entities (the ‘absolute slicing hypothesis’, as

’993 Patent		
		formulated in (Goldsmith, 1976, pp. 16-17)) has proved to be untenable. In non-linear, autosegmental phonology (Goldsmith, 1976, 1990) an analysis has been proposed in which different features are placed on different tiers. The various tiers represent the parallel activities of the articulators in speech, which do not necessarily begin and end simultaneously. In turn the tiers are connected by association lines. In this way, it is possible to indicate that the mapping between tiers is not always one to one. Assimilation phenomena can then be represented by the spreading of one feature from one segment to the adjacent one. On the basis of this theory, Li Deng and his colleagues have built ASR systems with which promising results have been obtained (Deng and Sun, 1994).” Strik, at 242.
Claim 19		
19	The computer-readable storage device of claim 17, wherein the language model is generated by modeling pronunciation dependencies across word boundaries.	As discussed above with respect to claim 17, Strik, either by itself, or in combination with Steinbiss, discloses, expressly or inherently, the computer-readable storage device of claim 17. *See supra* claim [17], which is incorporated by reference herein. Strik discloses, expressly or inherently, a “language model [that] is generated by modeling pronunciation dependencies across word boundaries.” Strik describes using “cross-word processes” in the language model as part of an ASR system. *See, e.g.*, “The majority of the contributions are concerned with variation at the segmental level. A common way of describing segmental pronunciation variation in the context of ASR is by indicating whether it refers to word-internal or to cross-word processes, because this choice is strongly related to the properties of the speech recognizer being used. As a matter of fact, the choice for word-internal variation, cross-word variation or both, is

’993 Patent		
		determined by factors such as the type of ASR, the language, and the level at which modeling will take place.” Strik, at 229. “Besides within-word variation, cross-word variation also occurs, especially in continuous speech. Therefore, cross-word variation should also be accounted for. A sort of compromise solution between the ease of modeling at the level of the lexicon and the need to model cross-word variation is to use multi-words (Beulen et al., 1998; Finke and Waibel, 1997; Kessens et al., 1999; Nock and Young, 1998; Pousse and Perennou, 1997; Ravishankar and Eskenazi, 1997; Riley et al., 1998; Sloboda and Waibel, 1996; Wester et al., 1998a). In this approach, sequences of words (usually called multi-words) are treated as one entity in the lexicon (see also Section 2.4.1) and the variations that result when the words are strung together are modeled by including different variants of the multi-words. It is important to note that, in general, with this approach only a small portion of cross-word variation is modeled, e.g., that occurring between words that figure in very frequent sequences. Besides the multi-word approach, other methods have been proposed to model cross-word variation such as (Aubert and Dugast, 1995; Blackburn and Young, 1995, 1996; Cohen and Mercer, 1975; Cremelie and Martens, 1995, 1997, 1998, 1999; Mercer and Cohen, 1987; Perennou and Brieussel-Pousse, 1998; Pousse and Perennou, 1997; Safra et al., 1998; Schiel et al., 1998; Wiseman and Downey, 1998).” Strik, at 229. “Given that both within-word and cross-word variation occur in running speech, it will probably be necessary to model both of them. This has already been done in (Beulen et al., 1998; Blackburn and Young, 1995, 1996; Cohen and Mercer, 1975; Cremelie and Martens, 1995, 1997, 1998, 1999; Finke and Waibel, 1997; Kessens et al., 1999; Mercer and Cohen,

’993 Patent		
		1987; Riley et al., 1998, 1999; Schiel et al., 1998; Sloboda and Waibel, 1996; Wester et al., 1998a).” Strik, at 229-30. “As was mentioned earlier, multi-words can also be added to the lexicon, in an attempt to model cross-word variation at the level of the lexicon. Optionally, the pronunciation variants of multi-words can also be included in the lexicon. By using multi-words Beulen et al. (1998) and Wester et al. (Kessens et al., 1999; Wester et al., 1998a) achieve a substantial improvement, while for Nock and Young (1998) this was not the case.” Strik, at 234. “Another component in which pronunciation variation can be taken into account is the language model (LM) (Cremelie and Martens, 1995, 1997, 1998, 1999; Deshmukh et al., 1996; Finke and Waibel, 1997; Fukada et al., 1998, 1999; Kessens et al., 1999; Lehtinen and Safra, 1998; Perennou and Brieussel-Pousse, 1998; Pousse and Perennou, 1997; Schiel et al., 1998; Wester et al., 1998a; Zeppenfeld et al., 1997). This can be done in several ways, as will be discussed below.” Strik, at 236. “*Method 1*. The easiest solution is to simply add the variants to the lexicon, and not to change the LMs at all. In this case, for every variant the probabilities for the word it belongs to are used. Since the statistics for the variants are not used, it is obvious that this is a sub-optimal solution. In the following two methods the statistics for the variants are employed.” Strik, at 236.

’993 Patent		
		“*Method 2*. The second solution is to use the variants themselves (instead of the underlying words) to calculate the N-grams (Kessens et al., 1999; Schiel et al., 1998; Wester et al., 1998a). For this procedure, a transcribed corpus is needed which contains information about the realized pronunciation variants. These transcriptions can be obtained in various ways, as has been discussed in Sections 2.2 and 2.4. The goal of this method is to find the string of variants V which maximizes $P(X\|V) * P(V)$.” Strik, at 236-37. “Another important difference between the two methods is that in the third method the context-dependence of pronunciation variants is not modeled directly in the LM. This can be a disadvantage as pronunciation variation is often context-dependent, e.g., liaison in French (Perennou and Brieussel-Pousse, 1998; Pousse and Perennou, 1997). Within the third method, this deficiency can be overcome by using classes of words instead of the words themselves, i.e., the classes of words that do or do not allow liaison (Perennou and Brieussel-Pousse, 1998; Pousse and Perennou, 1997). The probability of a pronunciation variant for a certain class is then represented in $P(V\|W)$, while the probability of sequences of word classes is stored in $P(W)$. Strik, at 237.